ChatGPT

HP ZGX Fury G1n AI Stationと知財業務への影響分析

Executive Summary

2026年10月26日に日本HPが発表し、10月27日に国内販売を開始した「HP ZGX Fury G1n AI Station」は、従来の「高性能ワークステーション」というより、**データセンター向けAI基盤を部門単位で保有できる“AIアプライアンス”に近い製品**である。日本での希望販売価格は税込2,849万円から。NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip、748GBのCPU/GPUコヒーレントメモリ、最大20 PFLOPSのFP4 AI性能を備え、HPは最大1兆パラメーター級モデルのローカル推論と、1,000億パラメーター超モデルのファインチューニングを訴求している。 [1]

MONOistの記事が重要なのは、単に「大きなモデルがデスクサイドで動く」点ではない。従来、企業の知財部門が高性能生成AIを使う際には、①機密発明情報をクラウドへ送る、②中央IT部門のGPU基盤を利用する、③小型のローカルモデルに性能を妥協する、という三択になりがちだった。ZGX Furyはその間に、**「クラウドに出したくない機密情報を、十分に大きなモデルで、部門内に閉じて処理する」第四の選択肢**を作る。HP自身もクラウドやデータセンターの全面代替ではなく、データの機密性、容量、利用頻度に応じたハイブリッド運用を想定している。 [2]

知財部門10〜50名という前提では、特に適合度が高いのは、**未公開発明・出願前資料を含むRAG、特許調査の意味検索／再ランキング、明細書・OA対応のドラフト支援、NDA・共同研究・ライセンス契約レビュー、侵害分析のclaim chart補助、社内技術文書を横断するIPランドスケープ**である。一方、「最大1兆パラメーターが動く」というモデル容量と「データセンター級の処理量」は同義ではない。ZGXの252GB HBM3eは7.1TB/sだが、残る496GBのLPDDR5Xは396GB/sであり、巨大モデルがコヒーレントメモリに収まっても、72〜8GPUのデータセンターシステムと同じ同時ユーザー数やトークン処理量が得られるわけではない。 [3]

同世代のNVIDIA DGX B300は8基のB300 GPU、2.3TBのGPUメモリ、144 PFLOPSのFP4推論性能、14.5kWの消費電力を持ち、GB300 NVL72は72基のBlackwell Ultra GPU、20TBのGPUメモリ、最大1,440 PFLOPS（sparse FP4）、最大142kW級のラック電力を必要とする。したがってZGX Furyの本質的な価値は「データセンターと同じ絶対性能」ではなく、**データセンター技術の一部を1.6kW、5U級、2,849万円からの部門設備へ圧縮したこと**にある。 [4]

知財部門への推奨は、**全面クラウド移行でも全面オンプレ移行でもなく、「Confidential-by-Default」のハイブリッド構成**である。未公開発明、研究ノート、ライセンス契約、侵害分析、訴訟準備資料などはローカルAI enclaveへ置き、公開特許の大規模再処理、短期のファインチューニング、ピーク負荷、最先端モデル比較だけを承認済みクラウドへバーストさせる。この方式は、HP自身が想定するハイブリッド像とも整合する。 [2]

ただし、ローカルAIは「安全なAI」と同義ではない。プロンプトインジェクション、機密情報の出力、モデル・コンテナのサプライチェーン、AIエージェントの過剰な権限などは残る。OWASPもRAGやファインチューニングだけではプロンプトインジェクションを完全に防げないと整理しており、NISTは生成AIをライフサイクル全体でGovern／Map／Measure／Manageすることを推奨している。 [5]

結論として、ZGX Furyは知財部門にとって「PCの更新案件」ではなく、「社内に小型の主権型AI基盤を持つか」という情報ガバナンス案件である。 2026〜2027年は1台でのPoCと評価基盤整備、2027〜2029年はRAGと限定的AIエージェントの本番化、2029〜2031年は複数ノード／次世代機への更新と、知財ワークフローそのもののAIネイティブ化が現実的なロードマップになる。

製品と技術仕様

日本HPの発表日は**2026年10月26日**、国内販売開始日は**2026年10月27日**である。希望販売価格は**28,490,000円（税込）から**で、税抜換算では約2,590万円となる。MONOistの記事も同日11時15分に公開されており、日本HPの発表内容に加えて、日本国内での設置、保守、100V/200V時の電力、想定販売規模などを取材している。 6

項目	HP ZGX Fury G1n AI Station	知財用途から見た意味
CPU	NVIDIA Grace、72-core Arm Neoverse V2	大量文書の前処理、索引、RAGパイプラインをGPUと近接して処理可能。 3
GPU	NVIDIA Blackwell Ultra ×1。FP4 Tensor Core最大20 PFLOPS（sparsity）、15 PFLOPS（非sparse）	大規模LLM、embedding、reranker、OCR後処理、マルチモデル推論を1台へ集約しやすい。 3
GPUメモリ	252GB HBM3e ECC、7.1TB/s	中〜大規模モデルをGPU内に置きやすい。 3
CPUメモリ	496GB LPDDR5X ECC、396GB/s	GPUメモリと合わせ748GBのコヒーレントメモリ。巨大モデルを収容できる一方、HBMより大幅に帯域が低い点は重要。 3
CPU-GPU接続	NVLink-C2C 900GB/s	従来のPCIe型CPU/GPU構成より巨大モデルのメモリ移動に有利。 3
モデル規模	最大約1兆パラメーター推論、1000億パラメーター超のファインチューニングをHPが訴求	「入る」ことと「高速・多数ユーザーで動く」ことを分けて評価する必要がある。 1
追加GPU	日本HP/MONOistではRTX PRO 6000 Blackwell Max-Q Workstation Edition 96GBをオプションとして説明	可視化、画像・CAD系AIなどをGB300と分業できる。 7
ストレージ	HP仕様ではNVMe M.2、SED対応。OS用とデータ用スロットを持ち、HPの構成情報では2TB/4TB級を選択可能	特許全文コーパスやembeddingは急速に容量を消費するため、実運用ではNAS/object storageとの併用が現実的。 8
ネットワーク	ConnectX-8、QSFP112 ×2、各400Gb/s、合計最大800Gb/s。加えて10GbEとBMC用1GbE	複数ノード化や高速ストレージ接続が可能。管理ネットワークはAIデータネットワークと分離すべき。 9
電源	最大1,600W。日本の100V系では利用電力に制約があり、HP/MONOistは100Vで1300W、200Vで1600Wと説明	フル性能を狙うなら200V系統を前提に設備確認すべき。100〜127V環境では専用20A回路が必要とHPガイドも明記。 10
冷却	内部液冷＋ラジエーター／ファンによる排熱。外部のデータセンター用冷水設備は不要	「机の横に置ける」物理サイズでも、騒音・熱負荷から日本HP取材ではサーバールーム等が現実的とされる。 11
ラック	5Uラックマウント対応、レールキット同梱	知財部門の席ではなく、施錠されたIT設備区画に置きやすい。 7

項目	HP ZGX Fury G1n AI Station	知財用途から見た意味
OS	出荷時Ubuntu 24.04 LTS＋NVIDIA AI Developer Tools。RHEL／Red Hat AI Enterprise対応・認証。Windows対応は日本HP発表時点で2026年12月以降予定	2026年10月27日時点ではLinux前提で評価するのが安全。Windows計画を前提に購入判断しない方がよい。 1
AIソフト	HP ZGX Toolkit、HP Z Runtime。CUDA-X系、コンテナ環境等。NVIDIA AI Enterpriseは別途企業向けライセンス	HPツールが無償でも、モデル・OS・商用AIミドルウェア・データのライセンスまで無償になるわけではない。 1
AIエージェント	NVIDIA OpenShell対応	agentをサンドボックス化し、ネットワークや実行場所を制限する構成が可能。 12
セキュリティ	TPM 2.0、UEFI 2.9、UEFI Secure Boot、SED、BMC。DGX Station基盤ではhardware root of trust、Redfish等も提供	部門AI基盤として必要なハードウェア信頼基盤と遠隔管理機能を持つ。 13
保証	日本向けは3年間オンサイト保証。MONOistによれば国内修理／交換体制を予定	ただし水冷機構等の故障時には本体交換を基本とするため、ミッションクリティカル用途は2台化またはクラウド退避経路が望ましい。 7

特に注意すべきなのが**748GBという数字の読み方**である。252GBのGPU HBM3eは7.1TB/sだが、496GBのCPU LPDDR5Xは396GB/sで、両者には約18倍の帯域差がある。したがって「1兆パラメーターのモデルがメモリ空間に収容できる」という能力は、72 GPUを高帯域NVLinkで接続したNVL72と同等の推論スループットを意味しない。知財用途では、1兆パラメーターモデルを常用することより、70B〜数百B級モデル、embedding、reranker、RAG、複数ユーザーを一台で同時に扱える余裕の方が実務価値につながりやすい。前半はNVIDIAの公開仕様から確認でき、後半はそこからの実務上の推論である。 3

またNVIDIA DGX Station基盤はMIGを最大7インスタンスに分割し、複数ユーザーへ隔離されたGPU資源を割り当てる機能を掲げている。10〜50名の知財部であれば、「各担当者にGPUワークステーションを配る」のではなく、ZGXを**部内共有AIサーバー**として使う方向が合理的である。 3

主要一次情報への直接リンクは、日本HP 2026年10月26日プレスリリース、HP ZGX Fury公式製品ページ、HP QuickSpecs、HP ZGX Fury User Guide、NVIDIA DGX Station公式ページ で確認できる。記事元はMONOist 2026年10月26日記事である。 14

データセンター級AIとの比較

ZGX Furyを評価する際には、「ワークステーション対普通のGPUサーバー」ではなく、NVIDIAの同世代Blackwell Ultraアーキテクチャと比べた方が位置付けが明確になる。

指標	HP ZGX Fury / DGX Station級	NVIDIA DGX B300	NVIDIA GB300 NVL72
GPU	Blackwell Ultra ×1	B300 Blackwell Ultra ×8	Blackwell Ultra ×72
CPU	Grace 72-core ×1	Xeon Platinum 6776P ×2	Grace CPU ×36、計2,592 Arm cores

指標	HP ZGX Fury / DGX Station級	NVIDIA DGX B300	NVIDIA GB300 NVL72
GPUメモリ	252GB HBM3e	約2.3TB	20TB
CPUメモリ	496GB LPDDR5X	2TB標準、最大4TB	17TB LPDDR5X
高速／コヒーレント系メモリ	合計748GB	GPU 2.3TB＋host memory	37TB fast memory
FP4 Tensor性能	20 PFLOPS sparse / 15 dense	144 PFLOPS	1,440 PFLOPS sparse / 1,080 dense
GPU間／CPU-GPU	NVLink-C2C 900GB/s	NVSwitch／NVLink	NVLink 130TB/s aggregate
AIネットワーク	ConnectX-8 最大800Gb/s	8×800Gb/s cluster network＋BlueField	各GPU系ConnectX-8、ラックスケールfabric
電力	最大1.6kW	14.5kW	ラック最大約142kW
物理形態	デスクサイド／5U	10RU	1フルラック、液冷
冷却	本体内液冷＋空冷排熱	データセンター設備前提	direct liquid cooling前提
適切な運用単位	部門／研究グループ	社内AI基盤	AI factory／大規模DC
国内HP価格	2,849万円〜	要見積	要見積

ZGX Furyの数値はHP/NVIDIA、DGX B300はNVIDIAのユーザーガイド、NVL72はNVIDIAの現行製品仕様・Enterprise Reference Architectureに基づく。DGX B300は14.5kWで8基のB300と2.3TB GPUメモリ、144 PFLOPS FP4推論性能を持つ。GB300 NVL72は72 GPU、20TB GPUメモリ、130TB/s NVLink、最大1,440 PFLOPS、ラック電力最大142kW級である。 15

単純なFP4ピーク値では、DGX B300はZGX級の約7.2倍、GB300 NVL72は約72倍である。興味深いのは後者がほぼGPU数比と一致する点だが、実アプリケーション性能はモデル構造、量子化、batch size、context length、KV cache、通信、ソフトウェア最適化に依存するため、この比率をそのまま「特許調査が72倍速い」と解釈してはいけない。数値の比率は公開仕様からの算術比較である。 16

一方、電力はZGXの1.6kWに対しDGX B300が14.5kW、GB300 NVL72はラック最大142kWである。ZGXは通常のPCよりは明らかに設備要件が重いものの、「受電容量、液冷CDU、冗長電源、高速fabricを持つAIデータセンターを新設する」こととは桁の違う設備案件である。 17

運用面ではこの差がさらに大きい。NVL72はGPUサーバーというより、電力、液冷、ネットワーク、OOB管理まで統合されたAI factory構成で、NVIDIAのReference Architectureは液漏れ検知、複数の33kW power shelf、OOBネットワーク等まで規定している。対してZGXはBMCや高速NICを持ちながら5Uで収まり、既存のサーバールームへ追加しやすい。 18

したがって、知財部門での比較軸は次のようになる。

観点	ZGX Fury	データセンターGPUクラスタ
巨大モデルを「動かせるか」	非常に強い	最も強い

観点	ZGX Fury	データセンターGPUクラスタ
同時利用・大量batch	中程度〜強	圧倒的に強い
大規模事前学習	不向き	本来の領域
LoRA/PEFT等の部門fine-tune	強い	強いが過剰設備になり得る
RAG/embedding/rerank	部門用途には非常に適合	全社規模なら適合
データを部外へ出さない	容易	自社DCなら容易
導入期間	比較的短い	DC設備次第で長い
スケール弾力性	低〜中	高い
障害時冗長性	単体では低い	構成次第で高い
IT運用難度	Linux/MLOps知識は必要	DC/HPC専門人員が必要
技術陳腐化リスク	1台単位なので限定	大規模CAPEXほど大きい

このため、ZGX Furyを「小型NVL72」と見るより、**“企業の機密データの隣に置ける、大容量メモリ型の共有AIノード”**と位置付ける方が適切である。HP自身もクラウド／データセンターの完全代替とはしていない。[2]

知財業務への導入シナリオとワークフロー変革

知財業務で最も大きく変わるのは、生成AIを「インターネット上のサービス」ではなく、**社内の検索・分析基盤の一部として扱えるようになること**である。

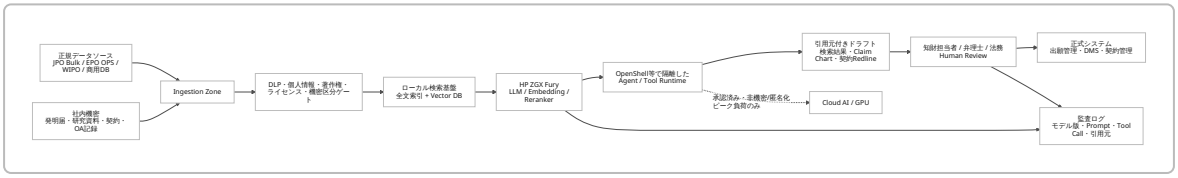
知財業務	ZGX導入後の典型処理	主な効果	主な制約・リスク
特許調査	公開公報→embedding→semantic search→cross-encoder rerank→LLM要約・対比	キーワード／分類検索だけでは拾いにくい概念類似を補助、長文読解を高速化	recall不足、分類／翻訳誤り、hallucination。最終検索をLLMだけに委ねない
出願支援	発明届・研究ノート→発明要素抽出→先行技術との差分→明細書構成・claim候補	弁理士・知財担当者の初稿作成時間を圧縮	未公開発明の漏えい、サポート不足、過度なclaim一般化、著作権・モデルライセンス
OA対応	引用文献とclaimを自動mapping→論点候補→応答案ドラフト	引用文献の一次読解を高速化	法的判断は人間必須、引用根拠の誤認
契約レビュー	NDA、共同研究、ライセンス、MTA等をRAG→条項抽出→社内playbookとの差分	大量契約のfirst-passレビューを自動化	governing law、例外条項、定義の連鎖を誤る危険
侵害調査	claim decomposition→製品資料・マニュアル・画像・Web資料→claim chart草案	claim chart作成を高速化	Web証拠の真正性、取得日時、chain-of-custody、誤った技術推定

知財業務	ZGX導入後の典型処理	主な効果	主な制約・リスク
IPランドスケープ	数百万文献のembedding、クラスタリング、企業・発明者・技術軸分析	R&D／M&A／競合分析を知財部内で反復可能	データ品質、family正規化、名寄せ、ライセンス
機密ナレッジQA	未公開出願、研究報告、社内審査記録、契約を閉域RAG	外部APIへ原文を送らずAI利用	内部者アクセス、ログへの機密残存、RAG権限漏れ
AIエージェント	文献収集、分類、要約、比較、報告書ドラフトをtool-callで連鎖	人が複数ツールを行き来する作業を削減	prompt injection、過剰権限、自動送信・自動出願等の事故

この中で最も重要なのは**特許データ取得方式**である。J-PlatPatは人が検索・閲覧する基本サービスとして提供されており、INPITは単純な大量ダウンロードやロボットによる定期収集を禁止し、プロフェッショナル用途には民間特許情報サービスを推奨している。さらに、明細書・請求項・図面等について著作権上の注意も明示されている。したがって「J-PlatPatをスクレイピングしてローカルRAGを作る」という設計は避けるべきである。 19

日本特許庁は別途、特許情報の一括ダウンロードサービスを提供しているため、ローカル検索基盤を構築するなら、JPOの正規バルクデータ、契約済み商用データ、許諾されたAPIを使うべきである。EPOもGUIへの自動検索ではなくRESTインターフェースの利用を求め、OPSには無料枠として週4GB等の条件を設けている。WIPO PATENTSCOPEにも自動クエリ・bulk取得等に利用条件がある。 20

したがって、推奨ワークフローは以下のようになる。



この構成では、LLMは「答えを決定する機械」ではなく、**検索候補の生成、順位付け、比較、要約、ドラフト作成を高速化する機械**として位置付ける。Stanford RegLabの法務AI研究では、RAGを採用した専門リーガルリサーチ製品でもhallucinationがなくなるわけではないことが実証されている。これは特許調査そのものを測定した研究ではないが、少なくとも「専門DB＋RAGなら無検証で法的結論を任せられる」という前提を否定する材料になる。 21

特許調査では特に、LLMによる意味検索は**recall向上の補助チャネル**にすべきであり、IPC/CPC/FI/F-term、キーワード、引用、family、出願人検索等を代替させるべきではない。INPIT自身もJ-PlatPatを基本的情報向けとし、高度なプロフェッショナル検索には専門サービスを推奨している。 19

侵害調査では完全air-gapにも弱点がある。競合製品の最新Web情報、マニュアル、ECサイト資料などは外部取得が必要だからである。そのため、ZGX自身に自由なWebアクセスを許可するより、**外部収集専用環境→マルウェア／prompt injection検査→証拠保全→一方向import**という構成が望ましい。これはOWASPが指摘する間接的prompt injectionやLLM supply-chainのリスクにも対応する設計である。 22

コスト試算とクラウド・オンプレ・ハイブリッド比較

以下は、知財部10〜50名、ZGX Fury 1台を共有AI基盤として3年間使うケースの**管理会計上の概算**である。HPの価格以外の導入工数、人件費、電力単価は企業ごとの差が極めて大きいため、本稿の比較用仮定として明示的に置いている。

基本ハードウェア価格は税込2,849万円からで、3年間のオンサイト保証が付く。 [6]

コスト項目	1台導入の試算	前提
ZGX Fury本体	**2,849万円〜**	HP国内希望販売価格、税込。税抜約2,590万円。 [12]
ラック・200V・UPS・配線・高速LAN	**50〜300万円**	本稿仮定。既存サーバールーム設備による
SSO、RBAC、SIEM、RAG、DMS連携、評価環境	**400〜1,200万円**	3〜8人月程度のSI/MLOps作業を想定した本稿仮定
初期導入合計	**約3,300〜4,350万円**	商用特許DB・モデルライセンスを除く
電力＋空調	**約20〜80万円/年**	平均0.6〜1.6kW、25〜35円/kWh、施設オーバーヘッド1.2〜1.5を仮定。最大システム電力は1.6kW。 [3]
内部運用人件費	**360〜1,440万円/年**	共通IT/MLOps 0.3〜0.8FTE、loaded cost 1,200〜1,800万円/FTE/年という本稿仮定
HP標準保守	**最初の3年は基本保証に内包**	国内3年間オンサイト保証。追加Care Pack等は別見積。 [7]
データ・モデル・ミドルウェア	**別枠**	HP ZGX Toolkitは無償だが、NVIDIA AI Enterprise、Red Hat、商用モデル、商用特許DB等は契約次第。 [1]
3年TCO目安	**約4,500〜9,200万円**	商用データ／モデル契約を除く本稿試算

この3年TCOで最も大きいのは電気ではなく**人件費とシステム統合費**である。1.6kW級の機械なので電力費はサーバールーム運営上無視できないものの、GPUデータセンターほど支配的ではない。むしろ「誰がモデルを更新するか」「誰が脆弱性を監視するか」「評価セットを維持するか」「アクセス権を管理するか」が継続費用になる。最大1.6kWという電力仕様はNVIDIA/HPの公表値に基づき、電力単価等は本稿の試算前提である。 [23]

クラウドとの比較には注意が要る。同世代の完全な1GPU GB300クラウドSKUとZGXを一対一比較する公開料金は限られるため、データセンター側の一例としてAWSの8×B300構成を参照する。AWS Capacity Blocksの2026年10月時点の公開ページではp6-b300.48xlargeの価格が地域・表示ページによりおおむね**US$112〜129/時間**程度で表示されている。 [24]

2026年10月26日の日本銀行公表のドル円中心相場157.20円/USDを換算基準とすると、約**1.77〜2.03万円/時間**に相当する。これは8 GPU B300なのでZGX 1台より計算資源が大幅に大きく、経済的な「同等性能比較」ではなく、あくまで同世代データセンター容量を借りる場合の予算規模を見るための参照値である。 [25]

AWS 8×B300利用量	年間GPU計算費の概算	コメント
200時間/年	約350〜410万円	sporadicなPoC、モデル評価向き
500時間/年	約880〜1,020万円	cloud burst向き
1,000時間/年	約1,770〜2,030万円	断続的だが相当量の開発
2,000時間/年	約3,530〜4,060万円	定常的利用になるとCAPEXとの比較が重要
24時間×365日	約1.55〜1.78億円/年	専用常時稼働ならオンプレ検討余地が大きい

上表にはストレージ、データ転送、private connectivity、バックアップ、ログ、運用人件費は含んでいない。またAWS Capacity Blocksは通常の従量VMとは予約・料金体系が異なり、料金自体も更新されるため、調達時点で再見積もりが必要である。 [26]

実務上の選択は次の通りになる。

比較軸	オンプレ ZGX	クラウドAI/GPU	ハイブリッド
初期CAPEX	高	低	中〜高
利用量増加時の限界費用	小	使用量に比例	中
稼働率が低い場合	不利	有利	有利
一定の高稼働	有利になりやすい	高額化しやすい	有利
機密／未公開発明	**最も扱いやすい**	契約・region・設定に依存	**非常に扱いやすい**
データ主権	強	provider／region依存	データ分類で制御
容量弾力性	低	**非常に高い**	**高い**
最新GPUへの追随	弱	**強い**	強い
BCP／冗長性	1台では弱	強い	**強い**
運用複雑度	中	中	**最も高い**
知財部10〜50名への適性	高	利用量が少ない場合高	**総合的に最も高い**

したがって、本稿では**ハイブリッドを第一推奨**とする。例えば平常時はZGXで機密RAG・推論を行い、年200〜500時間だけB300級クラウドを使うなら、クラウドGPU部分は概算年間350〜1,020万円程度で済む一方、大規模fine-tuneや一時的batch処理のために社内設備を過剰増強せずに済む。基礎料金はAWS公式公開価格と日銀為替からの本稿計算である。 [27]

なお、単純に「2,849万円÷クラウド時間単価」を行ってオンプレの損益分岐時間を求めるのは推奨しない。クラウドB300構成は8GPUでZGXより大幅に高性能であり、オンプレにはIT運用費、クラウドにはstorage/network/managed service費があるからである。**比較すべきKPIはGPU時間単価ではなく、「レビュー済み1案件当たり総コスト」「1,000件の特許を一次スクリーニングする総費用」「1契約のfirst-passにかかる費用」**である。性能仕様の差はNVIDIAの公開値から明確である。 [28]

セキュリティ・法務・知財保護

ZGXを知財部門へ導入する最大の動機は「データを外へ出さないこと」だが、**ローカル化はセキュリティ対策の一部であって、完了ではない**。

日本HPはTPM 2.0、UEFI Secure Boot、SED等を備え、NVIDIAのDGX Stationプラットフォームは BMC、Redfish、hardware root of trust等の企業向け管理・セキュリティ機能を備える。HPのユーザーガイドでは BMC自身にも設定、ログ、資格情報、service data等を格納するnonvolatile storageが存在し、廃棄・交換時にはBMC側のSecure Eraseも必要になることが示されている。つまり「SSDだけ消去して返却」では知財設備として不十分な場合がある。 [29]

推奨する技術構成は、ZGXの通常通信系、BMC/OOB管理系、ユーザー端末系、文書保管系を分割し、BMCを一般ユーザーLANから直接アクセスさせない構成である。またインターネットへの出口はdeny-by-defaultとし、モデル取得やアップデートはstaging環境で署名、hash、脆弱性、ライセンスを確認した後に昇格させるべきである。こうした対策は、OWASPが挙げるLLM supply-chain、prompt injection、sensitive information disclosureへの対策と整合する。 [30]

AIエージェントについてはさらに厳しくすべきである。OpenShellのようなsandboxを使い、ファイルシステム、外部ネットワーク、資格情報、tool-callをallowlist化し、**「読む」「検索する」「下書きを作る」と「送信する」「削除する」「出願する」「契約を承認する」を権限上分離**することが重要である。HPは OpenShellを分離されたsandboxでエージェントを実行する仕組みとして位置付けている。 [12]

ログも二面性を持つ。AI監査にはuser ID、案件ID、モデル名／hash、system prompt版、RAG corpus版、tool-call、参照文献、回答、human approvalを残したい。一方、promptや回答そのものには未公開発明や契約秘密が含まれるため、ログは新たな「最重要機密DB」になる。したがってSIEMへ全文を無条件転送せず、metadataとcontentを分離し、content logには案件単位RBAC、暗号化、retention、legal hold、削除ポリシーを設定するのが望ましい。NIST AI RMFもAIリスクをライフサイクル全体でGovern、Map、Measure、Manageする枠組みを提示している。 [31]

日本法上も、オンプレだから自由にデータを処理してよいわけではない。個人情報保護委員会は生成AI利用時の個人情報の適切な取扱いについて注意喚起を出しており、生成AIサービスへ個人データを入力する際には利用目的や、サービス側で応答以外の用途に利用されるかを確認する必要性を示している。ローカル運用は第三者AIサービスへの送信リスクを大幅に減らせるが、個人情報保護法上の安全管理や目的管理自体が消えるわけではない。 [32]

営業秘密についても同じである。経産省の2025年改訂「営業秘密管理指針」は、不正競争防止法による営業秘密保護を受けるための管理の考え方を示しており、2026年時点でも経産省が現行資料として掲載している。したがってZGXに格納しただけでは秘密管理性は自動的に成立せず、アクセス権、秘密表示、need-to-know、退職・異動時の権限剥奪などを継続する必要がある。 [33]

著作権については、「公開特許だからAI学習・再配布も完全に無制限」と考えない方がよい。INPITはJ-PlatPat上の明細書、請求項、要約、図面等について出願人等が著作権を有する場合があり、転載に許諾が必要になることがあると明示している。文化庁も生成AIと著作権について、開発・学習段階と生成・利用段階のリスクを区別したチェックリスト／ガイダンスを公表している。 [34]

また、**「HP ZGX Toolkitが無料」＝AIスタック全体が無料**ではない。HPが無料としているのはZGX Toolkit等であり、NVIDIAはDGX Station向けにNVIDIA AI Enterpriseを別途企業向けライセンスとして提供している。さらにLLM weight、embedding model、reranker、OCR、commercial patent corpus、翻訳エンジンにはそれぞれ別の利用条件があり得るため、モデル登録時に「名称、version、hash、取得元、license、

commercial-use可否、fine-tune可否、再配布可否」をmodel registryへ保持すべきである。前半はHP/NVIDIAの一次情報に基づき、後半はそこから導かれるライセンス管理上の推奨である。 [1]

海外拠点を持つ企業ではデータ主権も加わる。個人情報保護委員会がDeepSeekについて、データが中国所在サーバーに保存され中国法が適用される点を個別に注意喚起したことは、AIサービス選定において「モデル性能」だけでなく保存国・準拠法を確認する必要性を具体的に示している。ローカルAIは、この種のcross-border問題を減らす有力手段になる。 [35]

EUでAIを提供・展開する場合には、AI Actも確認が必要である。AI Actは段階的に適用され、2026年8月2日からは特定AIシステムやAI生成コンテンツに対する新しい透明性ルールが適用されている。内部の特許検索アシスタントが直ちにhigh-risk AIに該当するという意味ではないが、EU向け外部チャットボットや公開コンテンツ生成へ用途が拡張された場合には別途評価が必要である。 [36]

知財保護の実務としては、次の原則を採用することを推奨する。

領域	推奨ベストプラクティス
データ分類	Public / Internal / Confidential / Patent-pre-filing / Litigation等に分類し、AI利用許可を紐付ける
RAG	ユーザーのACLを検索時にも継承。「検索できない文書はLLMにも渡さない」
ネットワーク	user、AI data、storage、BMC/OOB、internet gatewayを分離
秘密情報	API keyやDB passwordをprompt/contextに埋め込まずvaultを使用
Agent	sandbox、deny-by-default、read-only初期設定、tool allowlist、人間承認
Model supply chain	hash固定、signed container、SBOM、model provenance、license registry
更新	新モデルを即本番投入せず、staging→評価→security review→承認→production
ログ	user/model/data/tool/source/human approvalを記録。ただしprompt全文の保管範囲は最小化
廃棄・修理	SSDだけでなくBMC、TPM、ローカルcache等のsanitizationを手順化
品質保証	patent search、claim chart、契約等ごとにgolden datasetを作り、更新ごとにregression test
人的統制	AIの結果には担当者／弁理士／法務の最終責任者を明示し、自動出願・自動契約承認は禁止

今後の影響予測と推奨アクション

2026年10月時点から3〜5年、すなわち**2029〜2031年**までを見ると、最も可能性が高い変化は「LLMが知財担当者を置き換える」ことではなく、**知財部門の標準業務が“人がDBを一件ずつ操作するプロセス”から、“AIが収集・比較・ドラフトし、人が検証・判断するプロセス”へ変わること**である。法務分野では専門RAG製品ですらhallucinationが残ることが実証されており、完全自律化よりhuman-in-the-loopの高度化の方が現実的である。 [21]

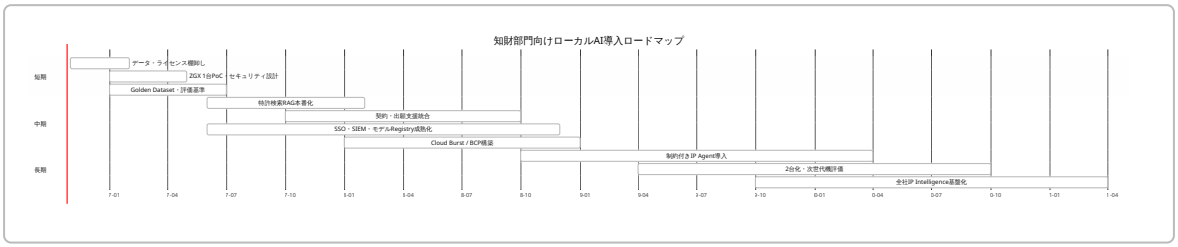
確度が高い予測は、機密性の高い部門ほどlocal／sovereign AIを採用すること、クラウドとのハイブリッドが主流になること、モデルそのものより「社内データ＋検索＋評価＋権限制御」が競争力になることであ

る。HP自身がプライバシー、データ主権、ローカル推論をZGX Furyの主要価値として訴求し、同時にクラウドの役割も否定していないことがこの方向を裏付ける。 6

確度が中〜高の予測として、特許調査は「検索式作成→結果閲覧」だけでなく、keyword/classification search、semantic retrieval、citation graph、LLM rerankingを並列に使う多段検索が標準化すると考えられる。WIPOが各国IP庁のAI導入事例を整理し、IP Australiaなどで自動検索支援が既に使われていることからも、特許検索へのAI組み込みは一時的流行ではない。 37

確度が中程度の予測として、知財部門専用のAI agentが発明届から先行技術収集、発明者への質問案、出願資料ドラフトまでを連鎖処理するようになる。ただしOpenShellのようなsandbox、OWASPが警告するexcessive agencyへの対策、human approval gateが重要になり、最終的な提出・外部送信まで無条件に自律化する企業は限定的と考えられる。 38

ロードマップは次のようにすることを推奨する。



短期、今後0〜6カ月では、まず購入より前に評価データを作ることが重要である。過去に調査結果が確定している20〜50件の先行技術調査、OA、契約、claim chartをgolden set化し、候補文献recall、引用正確率、レビュー時間、false negative、幻覚率、機密データ漏えいテストを測定する。同時にJPO/EPO/WIPO/商用DBのデータ利用条件を棚卸しし、「取得できる」データと「AI corpusとして保存・再利用できる」データを分ける。J-PlatPatの大量robotアクセスは禁止されているため、ここはPoC開始前に確定しておくべきである。 39

この段階では、いきなり「1兆パラメーターモデル」をKPIにしない方がよい。70B〜数百B級モデル、複数embedding、rerankerを比較し、**知財タスクの品質／1件当たり時間／消費電力／同時ユーザー数**で選ぶべきである。ZGXの748GBメモリはモデル選択の余裕を生むが、それ自体が業務品質を保証するわけではない。 3

中期、6〜24カ月では、特許検索と社内IPナレッジRAGを最初の本番対象にするのがよい。これらはAIが間違っても人が確認しやすく、検索候補や引用元を表示することで検証可能性を高く保てる。一方、自動出願、FTOの最終結論、契約承認、訴訟方針決定はhuman approvalを残す。Stanfordの法務AI研究が示すように、専門RAGでも信頼性問題は消えない。 40

利用者が20〜50名になり、ZGXが日常業務の中核になる段階では、単体障害が部門全体を止めるため、**2台目またはクラウドへのfailover**を検討すべきである。ZGX/DGX Stationは高速ConnectX-8を持ち、NVIDIAは複数Station接続や部門共有用途を想定している。 41

長期、24〜60カ月では、ハードウェアそのものを固定資産として守るより、ワークロードをcontainer、model registry、vector index、evaluation suiteとして可搬化することが重要になる。そうすれば次世代NVIDIA、クラウド、あるいは他社acceleratorへ移る際も、知財部門が築いたデータ・評価・ガバナンスを残せる。ZGXの価値はCUDA/NVIDIAデータセンター環境との連続性にある一方、その利便性は同時にvendor lock-inにもなり得るためである。NVIDIA自身もDGX Stationからクラウド／データセンターへのシームレスな展開を製品コンセプトとしている。 3

最終的には、知財部門のAI投資を「何人分の工数を削減したか」だけでは測らない方がよい。推奨KPIは、**既知先行技術の再発見率、候補文献recall、claim-to-evidenceの根拠付与率、契約上の重要条項の見逃し率、AI回答の引用検証合格率、レビュー時間、機密情報の外部送信件数、モデル更新後のregression数、案件当たり総コスト**である。NISTのAI RMFが重視する継続的な測定・管理とも整合する。 31

総合すると、HP ZGX Fury G1n AI Stationの知財分野へのインパクトは「より速い特許検索用PC」が登場したことではない。**これまでクラウド投入をためらっていた企業の最重要知財データに、70B〜1T級までの高度なAIを直接近づけられること**が本質である。そこから、未公開発明、契約、研究資料を含むRAG、部門専用AI、制約付きagentが現実的になる。その一方で、正規の特許データ取得、モデル・データライセンス、AIログの秘密管理、モデル更新時のregression test、human reviewを制度化できない組織では、2,849万円のGPUを導入しても知財業務の品質向上にはつながらない。HPの製品設計、INPIT/JPOのデータ利用条件、NIST/OWASPのAIリスク管理を総合すると、**2026〜2031年の勝負は計算性能そのものではなく、「機密データを安全にAIへ接続し、その出力を検証可能な知財ワークフローへ組み込めるか」に移る**と考えるのが妥当である。 42

---

1 6 12 14 38 42 日本HP、高度なAI開発に特化した高性能ワークステーション 「HP ZGX Fury G1n AI Station」を発表 | 日本HP
https://jp.ext.hp.com/info/newsroom/2026/20261026/

2 7 10 11 データセンター級のAIをローカルで、HPが大規模AI向けワークステーション発売：メカ設計ニュース - MONOist
https://monoist.itmedia.co.jp/mn/articles/2610/26/news027.html

3 9 15 23 41 Personal AI Supercomputer | NVIDIA DGX Station
https://www.nvidia.com/en-us/products/workstations/dgx-station/

4 16 17 28 Introduction to NVIDIA DGX B300 System
https://docs.nvidia.com/dgx/dgxb300-user-guide/introduction-to-dgxb300.html?utm_source=chatgpt.com

5 22 LLM01:2025 Prompt Injection - OWASP Gen AI Security Project
https://genai.owasp.org/llmrisk/llm01-prompt-injection/?utm_source=chatgpt.com

8 HP ZGX Fury: AI-Powered Workstation | HP® Official Site
https://www.hp.com/us-en/workstations/ai-stations/zgx-fury-ai-station.html

13 29 HP ZGX Fury G1n AI Station
https://h20195.www2.hp.com/v2/GetDocument.aspx?docname=c09316164&utm_source=chatgpt.com

18 System Hardware & Components — NVIDIA NVL72 AI ...
https://docs.nvidia.com/enterprise-reference-architectures/nvl72-ai-factory/latest/components.html?utm_source=chatgpt.com

19 34 39 ［INPIT］利用上のご案内
https://www.inpit.go.jp/j-platpat_info/guide/j-platpat_notice.html?utm_source=chatgpt.com

20 特許情報の一括ダウンロードサービスについて - 特許庁
https://www.jpo.go.jp/system/laws/sesaku/data/download.html?utm_source=chatgpt.com

21 40 Hallucination-Free? Assessing the Reliability of Leading AI ...
https://reglab.stanford.edu/publications/hallucination-free-assessing-the-reliability-of-leading-ai-legal-research-tools/?utm_source=chatgpt.com

24 26 27 Amazon EC2 Capacity Blocks for ML 요금
https://aws.amazon.com/ko/ec2/capacityblocks/pricing/?utm_source=chatgpt.com

25 boj.or.jp
https://www.boj.or.jp/statistics/market/forex/fxdaily/fxlist/fx261026.pdf

30 OWASP LLM02: 2025 Sensitive Information Disclosure
https://genai.owasp.org/llmrisk/llm022025-sensitive-information-disclosure/?utm_source=chatgpt.com

31 Artificial Intelligence Risk Management Framework
https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence?utm_source=chatgpt.com

32 生成AIサービスの利用に関する注意喚起等について
https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/?utm_source=chatgpt.com

33 営業秘密～営業秘密を守り活用する
https://www.meti.go.jp/policy/economy/chizai/chiteki/trade-secret.html?utm_source=chatgpt.com

35 DeepSeekに関する情報提供
https://www.ppc.go.jp/news/careful_information/250203_alert_deepseek/?utm_source=chatgpt.com

36 Safer and more transparent AI - European Commission
https://commission.europa.eu/news-and-media/news/safer-and-more-transparent-ai-2026-08-02_en?utm_source=chatgpt.com

37 Index of AI initiatives in IP offices
https://www.wipo.int/en/web/ai-tools-services/ipos-initiatives?utm_source=chatgpt.com