2026年は韓国産AI半導体飛躍の元年―「共同性能指標K-Perf、9.9兆ウォン予算で世界市場攻略」

科学技術情報通信部が主催する「2025AI半導体未来技術カンファレンス（AISFC 2025）」が今月10日、ソウルのロッテホテル・クリスタルボールルームで開催された。AISFCは2020年からAI半導体の技術動向とエコシステム活性化をテーマに開催されており、韓国内のAI半導体企業や学界の専門家が参加し、主要課題の議論とネットワーキングを行う場だ。今年は、韓国産AI半導体の性能を実需に即して評価する共同性能指標「K-Perf」協議体が正式に発足し、2026年AI半導体支援ロードマップが発表されるなど、AI三大強国入りに向けた基盤が示された。

「韓国、AI三大強国を目標に支援を3倍へ」

フュリオサAI（FuriosaAI）、リベリオンズ（Rebellions）、ハイパーアクセル（HyperAccel）など、韓国産AI半導体の性能を需要企業の要件に基づいて評価する共同性能指標K-Perf協議体が正式に発足した ／ 画像出典=FuriosaAI

ペ・ギョンフン副首相兼科学技術情報通信部長官は、「任期中に独自AIファウンデーションモデル、AIおよび量子技術、NPUを世界水準に引き上げる方策を導き出す。

韓国産AI半導体の性能はすでに成熟段階に入り、K-Perf宣言式がその出発点になる」と述べ、「AI三大強国となるには、韓国の研究者が同僚のように活用できるAIが必要であり、政府は分野別ファウンデーションモデルとサービス構築を進める」と語った。

ペ・ギョンフン副首相兼科学技術情報通信部長官がAISFC 2025でAI支援の成果と来年の目標を説明している ／ 画像出典=FuriosaAI

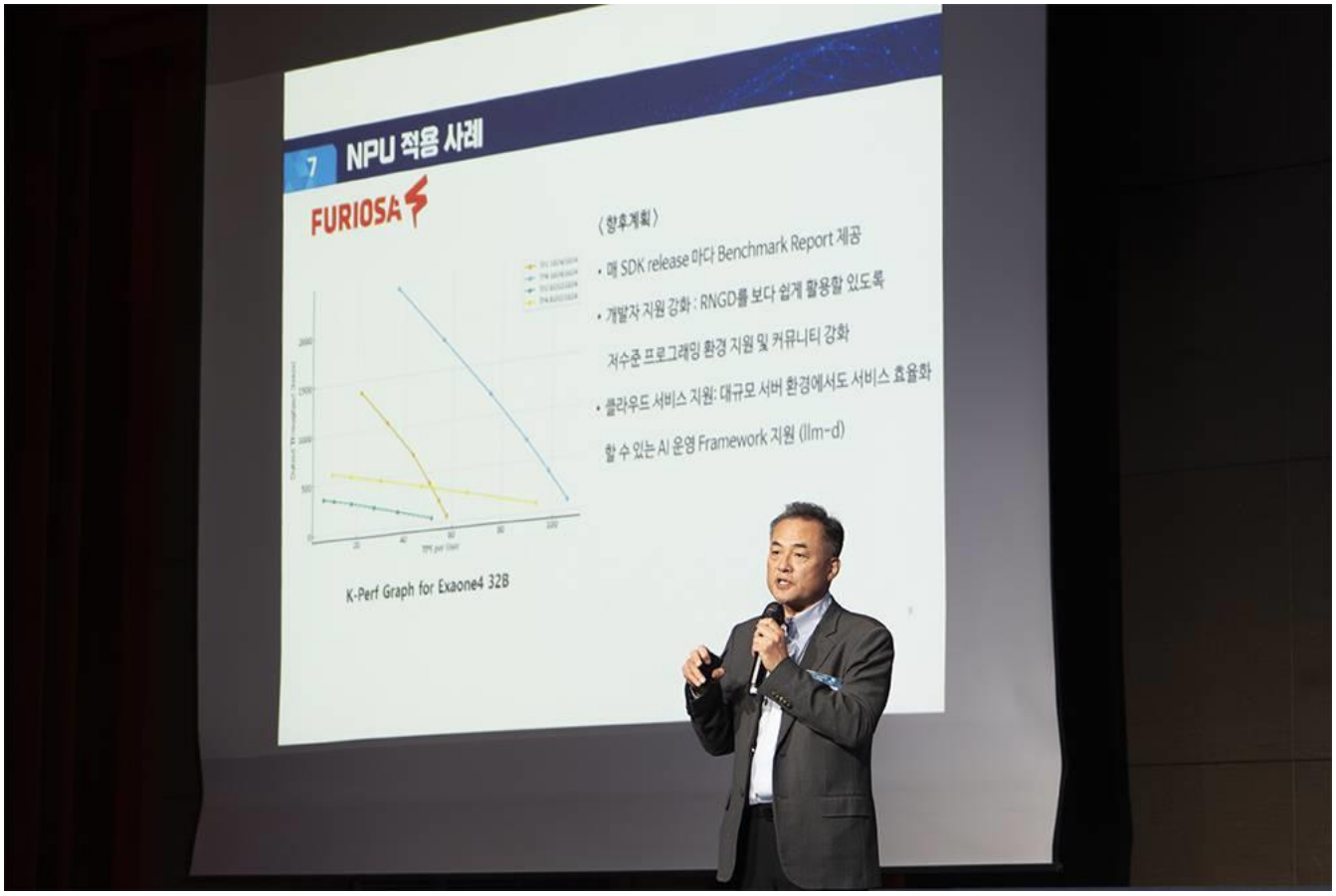
続けて「2025年が基盤構築の年であれば、2026年は本格的にAI強国、アジア太平洋地域のハブへと進む時点だ。政府は2026年に35兆ウォン規模のR&D予算を編成し、そのうちAI投資は従来比3倍となる9兆9000億ウォンとした。GoogleのTPUがNVIDIA GPUに匹敵する効率性を示したように、政府も持続的投資を通じてAIエコシステムの成長を支援する。韓国産AI半導体が第2のK-半導体成功の中核となるよう積極支援する」と強調した。

供給・需要が共同で評価する性能指標「K-Perf」正式始動

オ・ユンジェ情報通信企画評価院（IITP）半導体・量子PMがK-Perfの概要と評価指標を説明している ／ 画像出典=FuriosaAI

AISFC 2025の主要テーマの一つが、AI半導体性能を実使用に即して評価するK-Perfの発足だ。業界で広く使われるMLPerfは標準化されたワークロードを持つ一方、実運用性能との乖離や学習中心で推論評価が限定的という課題が指摘されてきた。これを受け、政府はAI半導体供給企業とクラウド・AI活用企業が協力する共同評価枠組みの構築に着手した。

供給側にはFuriosaAI、Rebellions、HyperAccelが参加し、需要側にはNaverクラウド、KTクラウド、NHNクラウド、サムスンSDS（Samsung SDS）、LG CNS、SKテレコム（SK Telecom）、LG AI研究院、カカオエンタープライズ（Kakao Enterprise）、モレー（Moreh）が名を連ねた。



K-Perfを用いてFuriosaAI、Rebellions、HyperAccelの半導体を評価した事例が公開された ／ 画像出典=FuriosaAI

主要テストは、Meta Llama 3.1（8B・405B）、Llama 3.3（70B）、EXAONE 4.0（32B）を活用し、今後はUpstageのWBLも追加予定だ。入力・出力長、同時ユーザー数、精度テスト、トークン処理速度、電力消費などを測定し、結果はExcelベースの測定表と2次元グラフで提示された。

オ・ユンジェPMは「需要側と供給側の性能認識のギャップは大きかった。K-Perfはその解消に向けた第一歩で、来年第1四半期に認証・検証手続きを構築し、将来的にはオンデバイスAIへも拡張する」と説明した。

K-Perf参加企業、2026年目標を提示

FuriosaAIのキム・ハンジュンCTOが第2世代半導体RNGDについて説明している ／画像出典=FuriosaAI

第3セッションでは、韓国のAI半導体企業および支援機関による発表が続いた。セッション3は「次世代AI半導体設計の高度化」をテーマに、FuriosaAIのキム・ハンジュンCTOによる「チップから市場へ、RNGDで実現するAI推論の効率化」、Rebellionsのオ・ジヌクCTOによる「効率的に駆動するフロンティアLLM・AIインフラの新時代」、HyperAccelのイ・ジンウォンCTOによる「持続可能なAIインフラとLLMサービスのためのLPU半導体」、DeepXのキム・ジョンウク副社長による「オンデバイスAI半導体、フィジカルAI時代への飛躍」がそれぞれ行われた。

FuriosaAIは2026年1月に第2世代NPU「RNGD」を商用化し、9月にはHBM3e 72GBを搭載したRNGD+、12月には2チップ構成のRNGD+ Max（HBM3e 144GB）を投入する計画だ。8枚のRNGDカードを搭載したサーバーは2026年3月に初公開し、2027年に第2世代サーバーを発売する。SDKは今月中にバージョン4.0を公開する。

SDK 4.0バージョンには、以下の機能が組み込まれている。▲ハイブリッド・バッチング（Hybrid Batching):異なる形態のAI推論リクエストを効率的にまとめて処理し、

NPU（Neural Processing Unit）の活用率とスループットを向上させる。▲プール・モデリング(Pooled Modeling): モデルの重み（ウェイト）をメモリに常時プール形式で保持し、すぐに再利用することで、最初の推論リクエスト時のローディング遅延を短縮する。▲NPUオペレーターのサポート拡大: NPUで直接実行可能なオペレーターのサポートを拡大。▲RNGDワークロード実行中にCPUやメモリの追加が必要となった際、Kubernetes（クバネティス）が動的に必要なリソースを割り当てる機能。▲PyTorchモデルを自動で最適化・コンパイルする**torch.compile()**のためのNPUバックエンドを提供する。SDK 4.0は、推論性能の最適化、AIモデルのメモリ使用・管理効率の改善、そしてNPUとKubernetesを中心としたインフラ統合を一層強化したバージョンと言える。

ペ・ギョンフン長官がFuriosaAIのRNGD半導体を視察している / 画像出典=FuriosaAI

ペ・ギョンフン長官もフュリオサAI、リベリオン、ハイパーアクセル、ディープエックス、モビリンツなどの主要AI半導体企業のブースを訪問し、製品の主要な仕様や事業化の現状について説明を受けた。

ペ・ギョンフン長官がFuriosaAIのRNGD半導体を視察している / 画像出典=FuriosaAI

FuriosaAIは、OpenAIの大規模オープンウェイト言語モデルであるgpt-oss-120Bを、2枚のRNGDカードで駆動するデモを披露した。gpt-oss-120Bは最低60GBのメモリを必要とし、1200億個のパラメータと128の専門家で構成されるMoE(Mixture of Experts)モデルだ。このモデルを約10ms水準の超低遅延環境で構築するには、NVIDIAのH100をマルチGPU構成で用いるか、MoEモデルの効率を大幅に高めたBlackwell B100などのチップが必要とされる。

FuriosaAIは、遅延を最小限に抑えるため、専門家ルーティングによって活性化され

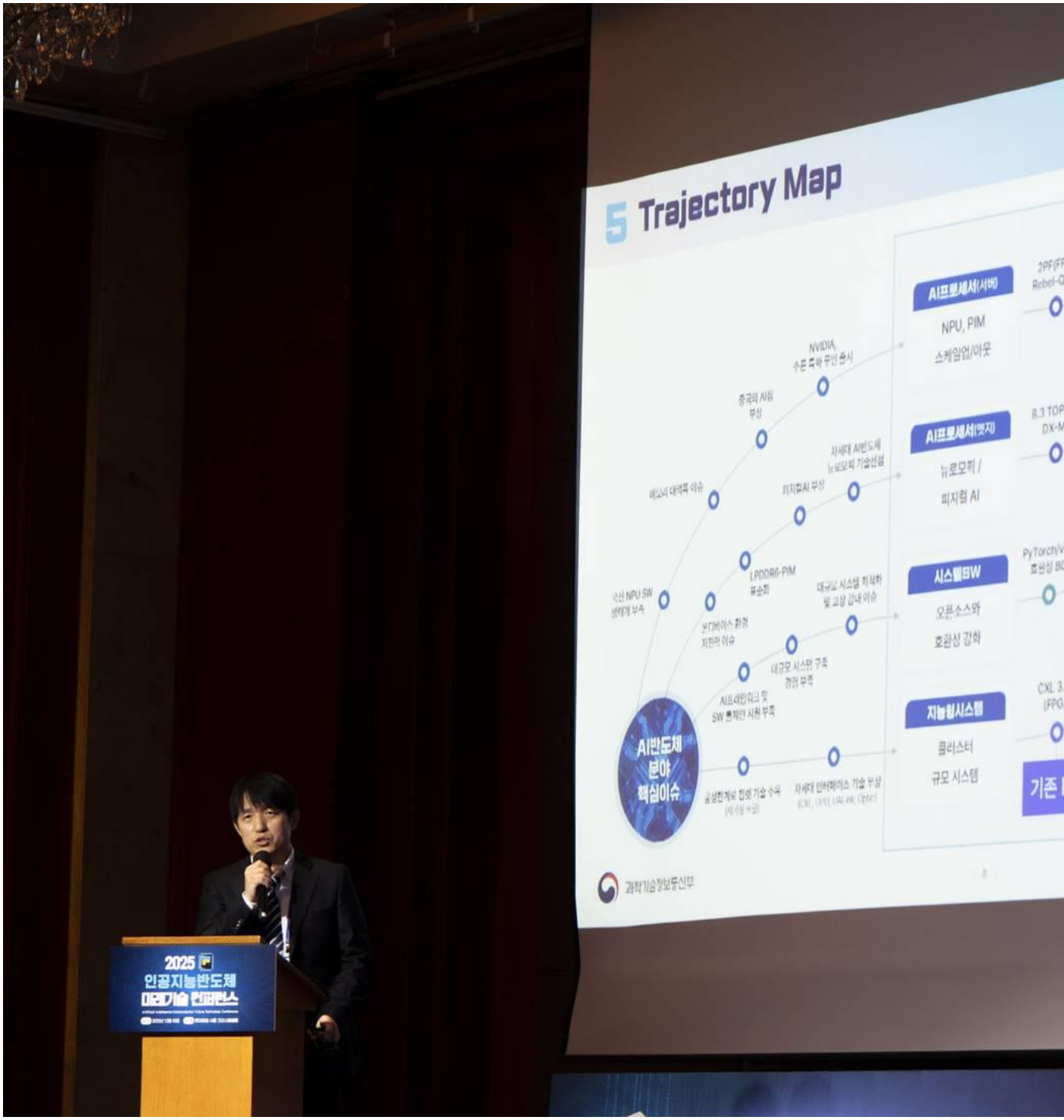
る一部の重みのみを選択的に計算するMoEの特性を積極的に活用した。また、gpt-oss-120Bが対応するMXFP4（4ビット混合精度量子化）フォーマットを、TCP（Tensor Reduction Processor）が直接演算できるよう最適化することで、メモリ帯域幅の使用を削減し、演算効率を大幅に向上させた。これら2つの最適化により、クエリ入力後約5.8msという超低遅延の応答速度を実現した。

IITP・NIPAが語る「2026年AI半導体支援事業」の方向性

情報通信企画評価院（IITP）のカン・ホソクチームリーダーが、2026 年半導体支援事業における重点推進方向の概要について説明した ／ 画像出典=FuriosaAI

第 4 セッションでは、情報通信企画評価院（IITP）が「2026 年 AI 半導体 R&D 事業の推進計画」および「AI 半導体実証を含む事業化支援の成果と計画」をテーマに発表を行った。IITP は科学技術情報通信部傘下で、ICT 分野における国家研究開発の企画・管理・評価を担う専門機関であり、技術事業化や専門人材育成において中核的な役割を果たしてい

る。韓国情報通信技術協会（TTA）、情報通信産業振興院（NIPA）とともに、K-Perf の試験認証、研究開発、実証・事業化を支援する主要機関の一つだ。



情報通信企画評価院は、プロセッシング・イン・メモリ（PIM）半導体の開発と、既存 AI 半導体に対するソフトウェア支援の拡大を 2026 年の主要目標に掲げている ／ 画像出典 =FuriosaAI

IITP のカン・ホソクチームリーダーは、「2025 年の支援事業は K-クラウドなどを通じて NPU 企業のスケール拡大に注力した。フィジカル AI 関連の予備妥当性事業としてオンデバイス AI を支援する事業も含まれ、PIM 半導体は 2021 年から新規事業として支援を継続している」と述べた。また、「2025 年の半導体性能基準は、オンデバイス AI 環境での 20B モデル駆動、ニューロモーフィック半導体では 1POPS（1 秒あたり 1000 兆回演算）性能、混合コンピューティングでは DNN-SNN（ディープニューラルネットワークースパイキングニューラルネットワーク）の混合対応を基準とした。さらに K-クラウドを通じた超巨大 AI モデル、光通信ベースのインターフェース支援も進めている」と説明した。

情報通信企画評価院は、2026 年には計 12 件の課題が運営され、2030 年を見据えて事業基盤を固める重要な年となると述べた ／ 画像出典=FuriosaAI

続けてカン・ホソク氏は、「2026 年の課題は計 12 件で、多くが予備妥当性調査段階にある。主要課題は LPDDR6-PIM ベースの AI アクセラレータと、それを活用するコントローラ開発で、ハードウェア開発およびソフトウェアフレームワーク開発も含まれる。また、供給企業と需要先のギャップを埋めるため、NVIDIA の NVLink のように AI 半導体チップ間通信ライブラリを最適化するシステムソフトウェアの方向性も検討している」と語った。PIM は INT8（8 ビット整数）基準で演算支援を確認中で、軽量 AI フレームワークは BF16（16 ビット浮動小数点）基準で進めているという。

さらに、NPU 企業が vLLM や PyTorch などのオープンソースフレームワーク対応を進めているものの、実際の採用につながっていない現状を踏まえ、互換性強化を目的とした競争型 R&D も推進する。競争課題は Meta Llama 8B が単一サーバーで安定して動作するかどうかを評価基準とし、各 NPU ハードウェアを活用する低レベル API の提供も求める予定だ。これは、AI 推論時にハードウェアアクセラレータの性能を最大限に引き出せるシステム基盤と完成度を確認する意図がある。

情報通信産業振興院(NIPA)は、韓国産 AI 半導体の需要拡大と実証事例の確保に注力している ／ 画像出典=FuriosaAI

NIPA は 2025 年の第 1 次補正予算を通じ、50TFLOPS 規模の AI 実証インフラ高度化、韓国産デバイス内 NPU 適用および AI 実証サービス構築、海外現地実証支援を実施した。続く第 2 次補正では、最新 AI モデルと韓国産 NPU の互換性確保、半導体設計用 IP 支援を行い、16 社・計 27 種の NPU 開発・高度化という成果を上げた。

NIPA のチョ・ジェホンチームリーダーは、「NIPA は研究開発、設計ソフトウェア支援、試作品検証から量産まで、AI 半導体生産プロセス全体を支援している。2025 年は 1 次・2 次補正を通じて普及基盤を構築し、27 の韓国産 AI 半導体が市場に登場、16 のファブレス企業が支援を受けた。500 万ドルを超える輸出成果を上げた。AI 企業と AI 半導体企業を連携した輸出支援も進めている」と述べた。2026 年事業は、完成した製品を基に需要創出、制度改善、人材雇用連携、海外進出までを包括的に支援する方針だ。

2025 年 AI 半導体事業は顕著な成果を上げ、2026 年が本格的なスタートとなる

イ・ジンウォン HyperAccel CTO（左）、キム・ハンジュン FuriosaAI CTO（中央）、オ・ジヌク Rebellions CTO（右）が、OpenAI の gpt-oss-120B を 2 枚の RNGD カードで駆動するデモを見て意見を交わしている ／ 画像出典=FuriosaAI

世界的に見て、AI 半導体を設計から生産まで一貫して行える国は、韓国、台湾、米国の 3 カ国のみだ。設計まで含めれば EU、日本、中国、インドも含まれるが、生産工程の制約を超えるのは容易ではない。中国は巨額の予算を投じて生産能力を高めているものの、5nm 以下の先端プロセスを量産できず競争で後れを取っている。一方、韓国はメモリ半導体とファウンドリーを強みに急速に競争力を高めており、現在の支援と挑戦が実を結べば、第 2 の半導体飛躍期を迎える可能性がある。

そうした中で、2025 年に 16 社から 27 種の NPU が登場したことは世界的にも注目すべき成果だ。ただし、現時点では売上面で大きな成果には至っておらず、各社はグローバル市

場での契約獲得に注力している。共同性能指標 K-Perf の登場は、国内需要企業の要件を反映した成功事例を創出し、その実績をもとに海外市場での契約につなげるための重要な試みといえる。K-Perf の成功と、2026 年における韓国産 AI 半導体の世界展開に期待が寄せられる。