



المقالة الاستراتيجية

التاريخ التحريري: 23 أكتوبر 2026

حين أصبحت الدمام مصنعًا للثواني

When Dammam Became a Factory for Seconds: Groq, Specialized Inference Silicon, and the Thermodynamics of Useful AI

في عرضٍ أمام جمهور سعودي، لم يكن الشيء اللافت أن نموذجًا لغويًا أجاب عن سؤال يتعلق برؤية 2030. تستطيع عشرات النماذج فعل ذلك. اللافت كان ما حاول جوناثان روس، مؤسس **Groq**، وطارق أمين، الذي كان يقود أرامكو الرقمية حينئذ، إظهاره في قمة الذكاء الاصطناعي العالمية عام 2024: تحويل المحادثة الصوتية مع آلة من تبادلٍ يقطع الإنسان إيقاعه انتظارًا للرد إلى حوار لا تبدو فيه الحوسبة عقبةً مرئية. تلك تجربة عرضتها الشركتان، لا اختبارًا مستقلاً منشوراً لزمّن الاستجابة؛ لكنها كانت كاشفة للسؤال الذي بُنيت حوله شركة كاملة: ماذا يحدث حين يصبح التأخر بين كلمتين مشكلةً في تصميم المعالج نفسه؟ [S01](#)

بعد شهر، أعلن خبر لا يشبه أخبار بناء مراكز البيانات المعتادة. تقول **Groq** إنها أدخلت عنقود استدلال في السعودية إلى الخدمة خلال ثمانية أيام في ديسمبر 2024؛ وفي فبراير 2025 قالت إن مركزها بالدمام صار يعالج طلبات عملاء من العالم عبر **GroqCloud**، تزامناً مع إعلان التزام استثماري سعودي بقيمة 1.5 مليار دولار للتوسع في بنيتها. هذه أرقام ووقائع أعلنتها الشركة عن نفسها؛ لا تعني أن المبلغ كله أنفق، ولا أن كل القدرة المخططة نُفّذت، ولا أنها تقدّم كشفاً محاسبياً للصادرات. لكنها تكفي لتبديل موضع المملكة على الخريطة التقنية: ليست كل إجابة تصل إلى الهاتف السعودي قد حُسبت خارجه؛ وقد توجد إجابة يطلبها مطوّر في بلد آخر وتُحسب داخل أرض سعودية. [S02](#)

هنا تنفتح مسألة أعمق من تجارة الخوادم. لماذا يكون الموقع الذي يحتفظ بوزن النموذج أقل أهمية أحياناً من الموقع الذي يستطيع إخراج الرمز التالي في جزء صغير من الثانية؟ ولماذا قررت شركة صنعت رقاقةً من أجل توليد الكلام أن تجعل الخليج محطةً في شبكة عالمية، ثم دخلت مع **NVIDIA**، التي تبدو منافسها الطبيعي، في علاقة ترخيص

وتقارب معماري؟ والأكثر إزعاجًا لجمال الأرقام: إذا صنعت المملكة ثواني أقل انتظارًا، فأين ذهبت الكهرباء التي أنفقتها، وأي ماء حمل الحرارة بعيدًا؟

هذه المقالة ليست عن الانقطاع ولا عن الحوكمة ولا عن «سيادة النموذج» بوصفها شعارًا. إنها عن صناعة زمن الاستدلال؛ وهي صناعة تبدأ في الصمت الكهربائي لذاكرة صغيرة داخل رقاقة، وتتمزّ بالترجم البرمجي (Compiler) والمُجدول والأنابيب الحرارية، وتنتهي بفاتورة موارد يمكن حسابها. وفي قلبها فرصة سعودية غير بديهية: تصدير أداءٍ زمني قابل للقياس، لا مجرد استضافة ملف اسمه نموذجًا.

ليست الإجابة كلامًا دفعة واحدة

لنفترض أنك سألت مساعدًا صوتيًا عن تقرير صناعي. ما يراه المستمع سلسلة جمل؛ ما تراه منظومة الاستدلال توالي رموز (tokens) تختارها شبكة عصبية احتماليًا. يدخل السؤال أولًا إلى مرحلة التمهيد (prefill): تُعالج رموز السياق بالتوازي نسبيًا، وتُنشأ تمثيلات داخلية وحالات مفاتيح وقيم الانتباه (key-value cache, KV cache). ثم تبدأ مرحلة التوليد (decode): يُنتج النموذج رمزًا، فيصبح ذلك الرمز جزءًا من السياق اللازم لإنتاج الذي يليه. ليست هذه قسمة أدبية للوقت؛ إنها قسمة فيزياء الحوسبة إلى عمليتين مختلفتين.

الـ Prefill كثيف بضرب المصفوفات والعمليات الحسابية الكبيرة. أما Decode لطلب واحد فينتظر المرور المتكرر على أوزان النموذج وحالات الانتباه؛ وقد يكون مقيدًا بنقل البيانات عبر الذاكرة أكثر من انشغاله بضرب الأرقام. ولهذا لا يكفي أن تقول إن رقاقة ما لديها «عدد أكبر من التريليونات من العمليات في الثانية». السؤال الأدق: ما مقدار العمل الحسابي الذي تنجزه مقابل كل بايت تُحضره من الذاكرة؟ يسمى مهندسو الأداء ذلك الكثافة الحسابية (arithmetic intensity)، ويستخدمون نموذج السقف (Roofline model) لفهم متى تختنق سرعة برنامج ما بقدرة الحساب ومتى تختنق بعرض حزمة الذاكرة.

يمكن زيادة إنتاج المصنع بإضافة خطوط عمل، ولكن هذا لا يجيب عن سؤال الزبون الذي ينتظر طلبه الآن. وعلى النحو نفسه، يختلف مجموع الرموز في الثانية (aggregate throughput) عن سرعة الرموز لمستخدم واحد (tokens per second per user). قد يُسعد مزوّد الخدمة رقمٌ هائل لأنه جمع آلاف الطلبات معًا، بينما يشعر مستخدمٌ واحد بأن النموذج يحادثه على مهل. وبين الشعورين تظهر أربعة مقاييس مختلفة: التأخر حتى أول رمز (TTFT)، وزمن الرمز الناتج (inter-token latency)، والانتظار في الطابور، والزمن من إرسال الطلب حتى آخر رمز (end-to-end latency). وتفصل وثائق Groq التشغيلية بينها بالفعل في مقاييسها القابلة للرصد؛ وهي تفصل أيضًا بين خدمات ذات أولوية زمنية وأعمال دفعية لا تنتظر تفاعلًا فورًا. S10؛ S09

خذ مثالًا حسابيًا لا علاقة له بسرعة رقاقة بعينها. إذا كان علينا توليد 180 رمزًا بسرعة 300 رمز/ثانية، فإن زمن التوليد النظري 0.6 ثانية. وإذا بلغت السرعة 900 رمز/ثانية، صار 0.2 ثانية؛ كسبنا 0.4 ثانية في تلك المرحلة فقط. لا يدخل في الحساب إرسال الكلام إلى النموذج، ولا تحويل الصوت، ولا قراءة أدوات خارجية، ولا الانتظار السابق

للتوليد. ولكن إذا كان وكيلٌ برمجي سيجري أربع جولات متتابة من التوليد بالمقادير نفسها، فقد تتراكم وفورات تلك المرحلة إلى 1.6 ثانية. في الأنظمة الوكيلية (*agentic systems*) يصبح أثر سرعة الرمز التتبعي مضاعفًا بتعدد الخطوات؛ لذلك يختلف اقتصاد الاستدلال عن اقتصاد تدريب النموذج مرة واحدة.

لماذا حملت Groq المُجدول إلى داخل الرقاقة؟

الرقاقة التي بنت عليها الشركة أطروحتها تسمى وحدة معالجة اللغة (*Language Processing Unit, LPU*). وقد يكون الاسم أقرب إلى تعريف تجاري لفئة حوسبة متخصصة منه إلى صنفٍ متفقي عليه في كتب المعمارية؛ لكن ما وراء الاسم هندسة تستحق التفكيك. إن أنظمة وحدات معالجة الرسومات (*Graphics Processing Units, GPU*) الحديثة قوية ومتعددة الأغراض. تستفيد من الذاكرة عالية عرض النطاق (*High Bandwidth Memory, HBM*) ومن وحدات الأنوية الموترية (*Tensor Cores*) ومجداول وتعليمات وطبقات برمجية تعالج طيفًا واسعًا من أعباء التعلم العميق. بالمقابل، اختارت Groq أن يكون ترتيب انتقال البيانات والعمليات إلى حد كبير معلومًا قبل أن يبدأ التنفيذ. S04؛ S05

في الرقاقة التقليدية يسأل البرنامج أثناء العمل عما أصبحت موارده جاهزة لتنفيذه. وفي تدفق البيانات المحدد مسبقًا (*statically scheduled dataflow*) يقوم المترجم البرمجي (*Compiler*) بتوزيع العمليات وحركة البيانات في جدول زمني، وتنتقل النتائج عبر معالجات وذاكرة ووصلات مرتبة كأنها خط إنتاج يعرف موعد وصول المادة إلى كل محطة. هذه الصورة تفسير للبنية المعلنة وليست ادعاء بأن كل ظروف الخدمة الخارجية تصبح حتمية: ما يزال الطابور والشبكة والتزامح بين المستخدمين ومتغيرات النماذج مصادر حقيقية للتفاوت. لكن جعل حركة البيانات داخل الآلة قابلة للتوقع يخفف عبئًا كان على العتاد أن يحسمه أثناء التنفيذ. S04

واختارت الشركة الذاكرة الساكنة داخل الشريحة (*Static Random-Access Memory, SRAM*) بدل الاعتماد الرئيسي على جلب الأوزان عبر ذاكرة عالية العرض خارج قالب المعالج. تحتفظ SRAM بالبتات باستخدام خلايا ترانزستورية سريعة ولا تحتاج إلى دورات تجديد شبيهة بذاكرة DRAM؛ إلا أن كل بت فيها أغلى في مساحة السيليكون. تشتري السرعة والكمون المنخفض على حساب كثافة التخزين. وهذه المفاضلة هي مغزى الرقاقة، لا شعار «ذاكرة أسرع عشر مرات» وحده. تعرض Groq رقمًا يزيد على 80 تيرابايت/ثانية لعرض حزمة SRAM في جيل من رقاقتها، وتقابله بنحو ثمانية تيرابايت/ثانية لبعض ترتيبات GPU/HBM. هذه مقارنة تسويقية مشروطة بتعريف المسار والجيل؛ ليست نتيجة تصلح للحكم على كل GPU، أو لتأكيد أن الاستدلال الكامل أسرع عشر مرات لكل نموذج. S04

تظهر الصعوبة الحقيقية في سؤال لطيف المظهر وقاسٍ هندسيًا: أين نضع الأوزان؟ نموذج افتراضي فيه 70 مليار مُعامل بدقة 16 بت يحتاج نحو 140 مليار بايت للأوزان وحدها، أي 140 جيجابايت عشريًا قبل حالات الانتباه والمساحات المؤقتة والتكرار والتنسيق بين الشرائح. وإذا استُخدمت ثمانية بتات للوزن، ينخفض الحجم الخام إلى 70 جيجابايت، لكن ذلك لا يجعل النشر مجانيًا؛ ثمة دقة رقمية ومقاييس تكميم وKV cache وخيارات توزيع. أعلنت NVIDIA في 2026 تصميم رُق Groq 3 LPX فيه 256 مسرعًا و128 جيجابايت SRAM إجمالًا، مع

عرض حزمة مععلن يبلغ 40 بيتابايت/ثانية للذاكرة داخل الرف. إذن حتى في جيل أحدث، قد تتجاوز الأوزان الخام ذات الـ 16 بت لطراز 70B مساحة SRAM داخل رف واحد، دون أن يعني هذا عجز النظام عن خدمته؛ فهناك تقسيم للأوزان وتكميم وذاكرة ومسرّعات ومكوّنات أخرى. [S08](#)

الاستنباط: كلما اقتربت البيانات من الحساب قل زمن جلبها، لكنك تصبح مطالبًا بأن تفكر في البرنامج بوصفه خريطة توزيع وزن وذاكرة وتواصل عبر رقائق. المعرفة الفنية التي يمكن تطويرها محليًا هنا لا تقف عند شراء الشريحة؛ إنها تبدأ من توزيع الأحمال المتوافق مع المترجم (Compiler-aware placement)، والتكميم الذي لا يفسد جودة التطبيق، وجدولة دفعات متغيرة الطول، وقياس التواتر الذي تنتقل به بيانات KV بين الموارد. لا حاجة إلى اختراع شريحة في يوم واحد كي تمتلك فرقًا بارعة في هذه الطبقات التشغيلية.

مقارنة مفاهيمية بين تدفق ذاكرة GPU وتدفق LPU

الشكل 1. رسم تفسيري مبتكر لا يمثل مخططًا تصنيعيًا أو قياسًا لأداء كلتا الرقائقتين. مصادر المبادئ: [S04](#)؛ [S07](#)؛ [S08](#).

ثلاث تجارب حقيقية، وثلاثة أنواع مختلفة من «النجاح»

التجربة الأولى، من منصة العرض إلى الدمام. في سبتمبر 2024 أعلنت أرامكو الرقمية و Groq تعاونًا على مركز استدلال كبير، مع تصوّر لربطه بسوق الخدمات الرقمية **nawat**. ظهرت محادثة صوتية على المسرح. هذا إثبات عرض لواجهة تفاعل، لا شهادة مستقلة بأن البنية المخطط لها كلها وجدت بالفعل. ثم جاء إعلان تشغيل عنقود في ديسمبر خلال ثمانية أيام، وإعلان تقديم GroqCloud من مركز الدمام في فبراير 2025. الفرق بين الثلاثة جذري: توقيع التعاون يُنشئ احتمالًا، والعرض يثبت سيناريو ضيقًا، أما الخدمة السحابية المُعلنة فتُشير إلى تشغيل يقدّم منتجًا للمطورين. ولا تبين الوثائق العامة المتاحة لنا جدول استهلاك الكهرباء أو استهلاك الماء أو عدد رموز الدمام وحدها. [S01](#)؛ [S02](#)

التجربة الثانية، النموذج المحلي لا يحتاج إلى رقاقة محلية الصنع حتى يشتغل هنا. عرضت Groq في LEAP 2025 تشغيل نموذج عَلام سعودي ونماذج استدلال وتوليد كلام بالإنجليزية والعربية. وفي مايو أصبحت، بحسب بيانها، مزودًا رسميًا للاستدلال لشركة **HUMAIN** السعودية. التجربة الأولى عرضت نماذج واستخدامات؛ والثانية علاقة تزويد معلنة. وهنا فصلٌ صناعي ينبغي ألا تضيق دفته بين الخطب: يمكن أن يكون النموذج مطوّرًا في السعودية، والمعالج أمريكي التصميم والتصنيع وفق إفصاح Groq، ومكان حساب الطلب سعوديًا. هذه ثلاث طبقات مختلفة من سلسلة القيمة، ولا تصبح واحدة بمجرد حملها الاسم الوطني نفسه. [S02](#)؛ [S03](#)

التجربة الثالثة، الخصم الذي أصبح جزءًا من الآلة نفسها. في ديسمبر 2025 أعلنت Groq اتفاق ترخيص غير حصري لتقنية الاستدلال مع NVIDIA، مع انتقال جوناثان روس وسني مادرا ومهندسين إلى NVIDIA، فيما بقيت Groq شركة مستقلة بحسب بيانها. في أغسطس 2026 أعلنت NVIDIA أن **Groq 3 LPX** دخل الإنتاج الكامل، وأن

Nebius أول مزود سحابي يتبناه؛ بينما أعلنت Groq أنها **تخطط** لأن تكون من أوائل من يضيفونه بالتعاون مع Dell. مهم ألا يُقرأ «دخل المنتج التصنيع» بوصفه «رُكّب في مركز سعودي»، فلا دليل منشورًا هنا على ذلك. [S06](#)؛ [S07](#)؛ [S12](#)

الأهم أن تصميم NVIDIA الجديد جمع عائلتين من الحوسبة: أجهزة **Vera Rubin NVL72** ذات GPU، مع رفوف LPX المتخصصة، وقد وصفت NVIDIA المكوّنين بأنهما يشاركان في حساب طبقات النموذج لخدمة سياقات طويلة وتوليد سريع. هذه ليست حكاية فوز رقاقة وإحالة الأخرى إلى التقاعد؛ إنها اعتراف معماري بأن **الاستدلال ليس عملية واحدة يجب أن يعالجها نوعٌ واحد من السيليكون بالضرورة**. قد تسند معالجة السياق الكثيف إلى نمط من الحوسبة، وتوليد الرموز المتتابة إلى نمط مختلف، وفق توزيع مفصل على مستوى العمل والشبكة والذاكرة. ونُشرت نتيجة مرجعية بلغت نحو **3400 رمز ناتج/ثانية** في اختبار **Gemma 4 31B** بسياق 100 ألف رمز وفق NVIDIA نقلًا عن Artificial Analysis؛ الرقم مربوط بذلك النموذج والاختبار والمنصة، وليس وعدًا لكل مستفيد ولا نتيجة خاصة بالدمام. [S07](#)؛ [S08](#)

ثم مفارقة جغرافية صغيرة تمنع من رسم خريطة متسعة: ما زالت بيانات 2025 تصف التشغيل من **الدمام**؛ بينما تعرض صفحة منصة Groq المعاصرة نقطة **SA-1 — Riyadh** ضمن شبكتها، وتقول إنها تشغل 13 مركزًا عالميًا. لا توضح هذه الصفحة إن كانت التسمية عنوانًا تشغيليًا مختلفًا، أو موقعًا جديدًا، أو تغييرًا في كيفية تصنيف نقاط الوصول. لا **نجمع الدمام والرياض حسابيًا بوصفهما مركزين منفصلين**، ولا **نفترض انتقالًا من إحداهما إلى الأخرى**. ما نعرفه أن السعودية باتت عقدة تشغيل في شبكة Groq المعلنة؛ وما لا نعرفه هو تركيبها الدقيق وتوزيع أحمالها الحالي. [S02](#)؛ [S08](#)

ليس كل رمز خرج من الشريحة وصل في الوقت المفيد

إذا أنتج معالج ألف رمز في الثانية ولكنه انتظر طويلًا في طابور الدخول، خسر التطبيق شيئًا لا يظهر في شريحة المواصفات. وإذا كان الوكيل يستدعي أربع أدوات عبر قارات متباعدة، فإن تسريع Decode لا يزيل زمن الشبكات أو تنفيذ الأداة. لهذا تصبح **خريطة زمن الطلب** أكثر فائدة من التفاخر بمعدل وحيد: زمن رحلة الشبكة (*round-trip time*)، زمن الطابور، زمن Prefill، الفاصل بين رموز Decode، انتظار الأدوات، ومعدل تحقق المهمة كاملة عند جودتها المطلوبة.

وثائق Groq التجارية نفسها تعطينا صورة عملية لا تتطلب اختراع تجربة على بنيتها: فيها فئات **performance** للأعمال ذات متطلبات استجابة مشددة، و**on_demand** للطلب المعتاد، و**flex** لأعمال تستبدل بعض ضمانات الخدمة بحدود معالجة أوسع؛ وهناك **Batch API** مستقل بمدد وحدود مختلفة. لا تعني هذه المزايا أن مُجدولًا حراريًا يعمل بالفعل في الشركة؛ لكنها تثبت وجود **فروق زمنية في قيمة العمل** يمكن أن تستوعبها خدمة استدلال واقعية. جواب المكالمات الهاتفية يجب ألا ينتظر نافذة جوٍ أبرد. أما تلخيص أرشيف ضخم أو إجراء مجموعة تقييمات للنماذج فقد يملك ساعات من المرونة، إذا سمح العقد وطبيعة البيانات بذلك. [S09](#)؛ [S10](#)

وهكذا يمكن تعريف سلعة جديدة لمن يدير السحابة السعودية: رمز مقبول الجودة سُلم داخل مهلة معلومة، مع حساب الطاقة والماء اللازمين لإنتاجه. عدد الرموز الخام وحده خادع؛ فقد ينتج النظام رموزًا طويلة لا يحتاجها أحد، أو يعيد إجابة فاشلة، أو يُكمل دفعة بعد موعدها. والطاقة التي تستهلكها شريحة في الخمول أو يعيد التبريد سحبها لا تظهر في عداد الرموز أصلًا. تحتاج الصناعة إلى ثلاث عدادات متقاطعة: عمل مفيد، زمن نافذ، ومورد مادي.

الطقس ليس خلفية المبنى؛ قد يصبح مُدخلًا للخوارزمية

في علوم الحاسب يُسمى وضع مهمة في فتحة وقتية مناسبة **جدولة (scheduling)**. ولكن ماذا لو أدخلنا الهواء الخارجي إلى دالة تكلفة تلك المهمة؟ ليست درجة الحرارة الظاهرة على هاتفك كافية. يفرق مهندس التبريد بين **حرارة البصيلة الجافة (dry-bulb temperature)** التي يشير إليها مقياس حرارة مألوف، و**حرارة البصيلة الرطبة (wet-bulb temperature)** التي تتعلق بإمكان التبخر عند الرطوبة المتاحة. الهواء الحار الجاف قد يتيح تبريدًا بالتبخير على نحو مختلف من الهواء الساحلي الرطب؛ والليل الساحلي ليس بالضرورة «ليلاً باردًا» من منظور قدرة الهواء على استقبال بخار إضافي.

في مركز استدلال لا تختفي طاقة الكهرباء. معظمها يتحول في النهاية إلى حرارة يجب نقلها بعيدًا عن المعالجات والذاكرة ومزودات الطاقة. ويمكن حمل الحرارة بسائل يمر فوق **ألواح تبريد (cold plates)**، ثم ينقلها عبر مبادلات حرارية ومضخات إلى منظومة رفض الحرارة خارج المنشأة. نقل الحرارة داخل حلقة مغلقة من السائل لا يعني أن **المركز لا يستهلك ماءً**، مثلما لا يعني أن كل مركز سائل التبريد يستخدم أبراج تبخير. يتحدد أثر الماء عند طريقة رفض الحرارة، ومياه التعويض، وأي معالجة أو تصريف، ومصدر الكهرباء أيضًا. S11؛ S13

هناك خيار **جاف** يرفض الحرارة إلى الهواء بوسائط ميكانيكية من دون استخدام ماء تبخيري كبير في موقع المنشأة، لكنه قد يحتاج كهرباء أكثر حين يكون الهواء الخارجي حارًا. وهناك خيار **تبخيري** قد يخفض طاقة التبريد الكهربائية في ظروف معينة مقابل استهلاك ماء في الموقع. لا توجد «طريقة رابحة» في كل ساعة وكل مكان. الدليل الذي نشرته وزارة الطاقة الأمريكية يعرّف **PUE** بوصفه نسبة طاقة المنشأة إلى طاقة المعدات المعلوماتية، و**WUE** للموقع بوصفه كمية ماء المنشأة لكل كيلوواط-ساعة من طاقة IT. لكن المؤشرين لا يبلغانك بمفردهما كم مهمة نافعة أنجزت ولا كم ماء استُهلك عند توليد الكهرباء خارج البوابة. S11

يمكن فهم الفيزياء بملاحظة صغيرة: يحتاج تبخير كيلوغرام ماء عند نحو 25 درجة مئوية إلى قرابة **2.44 ميغاجول** من الحرارة الكامنة. والكيلوواط-ساعة يساوي 3.6 ميغاجول. لو فُرض، على نحو مثالي مجرد، أن كل حرارة كيلوواط-ساعة تُزال بالتبخير، لأصبح التبخر النظري قريبًا من **1.47 لتر ماء**. هذا ليس رقم **WUE** لأي مركز، ولا ماء كل سؤال، ولا يشمل ماء التصريف والفاقد والسلسلة الكهربائية. إنه علاقة طاقة بانتقال طور مادة؛ مفيد لأنه يذكّرنا بأن البتات وإن بدت عديمة الوزن فحرارتها ليست كذلك. S14

تجربة قابلة للإعادة: هل ينقل المُجدول الحرارة من ساعة إلى أخرى؟

أجرينا لهذا المقال اختبارًا حسابيًا صغيرًا فعلي التنفيذ بالكامل، اصطناعي المدخلات بالكامل. بنينا نافذة من 48 ساعة وسلسلتين جويتين مُختلقتين: إحداهما «حارة وجافة» وأخرى «ساحلية رطبة». لم نسمّهما الرياض والدمام لأننا لم نحصل على أرصاد سعودية ساعية؛ فهي سيناريوهات رياضية معقولة الشكل لا بيانات طقس. أدخلنا في كل ساعة عملاً تفاعلياً يجب تنفيذه فوراً، ودفعاً مرنة يمكن تأخيرها حتى تسع ساعات؛ وحددنا سقفاً لقدرة IT، واستهلاكاً ثابتاً في الخمول، ودالتي كهرباء وماء للتبريد الجاف والتبخيري. جرى التحسين على مرحلتين منفصلتين: اختارت الأولى، لكل ساعة، وضع التبريد الأدنى كلفة وفق وزن الماء؛ ثم وُزعت البرمجة الخطية (*linear programming*) العمل الدفعي على الساعات مع الحفاظ على جميع المواعيد وعلى المقدار نفسه من حجم العمل الاصطناعي. لا يكافئ هذا اختياراً مشتركاً إلا لأن التكاليف خطية وقابلة للفصل، والتبديل بين الأوضاع مجاني وغير مقيد في النموذج؛ وهي افتراضات تحليلية لا وصفٌ واقعي للتحكم في منشأة تبريد. المعامل الذي يوازن لتر الماء مع كيلوواط-ساعة الكهرباء مجرد وزن رياضي، لا سعر سوق سعودي. كل فرضية ودالة ومعادلة في البحث المرافق وشيفرة التنفيذ. S17

في السيناريو الحار الجاف، كانت السياسة التي تبدأ بالأعمال الدفعية عند أول فرصة وتختار نمط التبريد الأقل كهرباءً تستهلك 5192.71 كيلوواط-ساعة و4829 لتر ماء في 48 ساعة. خفّض اختيار وضع التبريد في كل ساعة ثم تحسين توقيت الدفعات بهدف الكهرباء وحدها الطاقة إلى 5179.88 كيلوواط-ساعة، مع 4167 لترًا تقريبًا. وعندما أدخلنا وزنًا محدودًا للماء قدره 0.02 كيلوواط-ساعة لكل لتر، انتقل الحل إلى 5188.14 كيلوواط-ساعة و3559.52 لترًا. أي إن المقارنة مع حل الحد الأدنى للكهرباء تشتري توفير نحو 607 لترات بزيادة نحو 8.26 كيلوواط-ساعة. هذه ليست نسبة وفورات حقيقية قابلة للتعميم؛ إنها نتيجة مشروطة بوظائف تبريد صُممت هنا لأغراض البحث. S17

وفي السيناريو الساحلي الرطب صارت المفاضلة مختلفة: حل تقليل الطاقة احتاج 5245.08 كيلوواط-ساعة و4893.67 لترًا؛ ومع الوزن نفسه للماء أصبح 5263.52 كيلوواط-ساعة و3640.80 لترًا. تضاف نحو 18.43 كيلوواط-ساعة لتخفيض نحو 1253 لترًا في هذا العالم الاصطناعي. جميع الوظائف الدفعية الـ48 اكتملت ضمن نافذتها في السيناريوهين، وبقي حجم العمل الاصطناعي W نحو 4394.2 وحدة مطبّعة؛ وهذا ليس عدد المهام المقبولة Y . لهذا تُظهر الأرقام آليةً ممكنة لا مجرد لعبةٍ في القسمة: النتيجة ليست «الأبرد أفضل»؛ إنها موازنة وقت إتمام العمل مع التكلفة الحدية للحرارة والماء. S17

منحنى المفاضلة بين ماء التبريد وكهربائه في سيناريوهين اصطناعيين

الشكل 2. نتائج حل فعلي لبرنامج خطي على مدخلات مختلفة؛ كل نقطة تعالج العمل نفسه من دون تجاوز المواعيد. المحوران بوحدات النموذج الاصطناعي، وليسًا قياسًا لمنشأة سعودية.

ولأن العمل التفاعلي لا يتحرك في التجربة أصلًا، فالمنحنى لا يُشتري بإبطاء محادثة حيّة. الذي يتحرك فقط عمل له نافذة تسمح بالحركة. بهذا المعنى لا يكون المجدول الحراري زرًا عامًا لكل الطلبات، بل أداة تُعرف قدرتها من مساحة المرونة الزمنية المتاحة فعليًا. وإذا غابت هذه المساحة، فليس للبرنامج أن يتظاهر بأنه سيغيّر الطقس.

تحريك وحدات العمل الدفعي عبر اليوم الاصطناعي

الشكل 3. مثال توقيت العمل الدفعي في السيناريو الحار الجاف. الخطوط من التنفيذ المحلي؛ لا تمثل أحمال Groq أو HUMAIN الفعلية.

بين مُجدول المترجم البرمجي (Compiler) ومُجدول المدينة

هناك اتصال معرفي جميل بين عمل Groq داخل الشريحة وما نقترحه خارجها، ولكنه تشابه في منطق التنسيق لا تطابق في الوظيفة. في المستوى الأول، يرتب المترجم البرمجي (Compiler) انتقال نتائج الضرب والمصفوفات عبر وحدات الرقاقة؛ لا يستطيع الانتظار حتى تبرد مدينة كاملة قبل إنتاج الرمز التالي لمستخدم حي. وفي المستوى الثاني، يقرر مُجدول الخدمة أي دفعة تقييم أو معالجة وثائق تُوزع على أي ساعة وموقع وتكوين تبريد. الأول يعمل على مقياس نانوثانية أو دورات ساعة المعالج، والثاني على مقياس دقائق وساعات. يخضع كلاهما لسؤال واحد بصياغتين: ما العمل الذي يمكن نقله دون أن تُخرق شروط صحته وزمنه؟

والأدب الهندسي هنا أن نرفض المبالغة. أبحاث أخرى تعمل بالفعل على جدولة الحمل بحسب استهلاك الماء والضغط المحلي، منها نسخة arXiv لعام 2026 من **Balancing Bits and Drops**. وفي أغسطس قدّم باحثون PowerSlider على نظام SGLang لقيود القدرة الكهربائية اللحظية؛ شرحوا أن خفض تردد GPU يصيب Decode و Prefill بصورة غير متجانسة، ونشروا نسب Goodput تجريبية ضمن إعدادهم الخاص، لا ضمن مركز سعودي ولا على LPU. الجديد الذي نقترح اختباره هنا ليس اختراع «الجدولة البيئية» من العدم، بل وصل التبريد المتغير لحظيًا بمواصفات عبء الاستدلال الفعلية داخل منشآت سعودية، وربطه بمؤشر عمل نافذ الجودة، عبر قياسات قابلة للمقارنة. [S15](#)؛ [S16](#)

لو فكرنا كفريق تشغيل حقيقي، بدأ الأمر بجمع أثر واحد للطلب: حجم التمهيد، عدد رموز المخرجات، حجم KV cache، عدد خطوات الأدوات، زمن بداية الطلب وأول رمز، الحد الأعلى المقبول عند المستخدم، وكم مرة أعيدت المهمة. ثم نضيف من الطبقة الفيزيائية: طاقة PDU للرف، وحالة المضخات والمراوح، وكفاءة مبادل الحرارة، وحرارتي البصيلة الجافة والرطوبة، وعداد مياه التعويض، وحدود سريان التبريد. قد يكون تحسين استخدام النموذج عبر التكميم الدقيق أو تخزين بادئات مشتركة (*prefix caching*) أهم من نقل مهمة إلى المساء. وقد يكون تفريق مرحلتَي Decode و Prefill بين موارد متخصصة أولى من كل إجراء حراري. لا نستنتج ترتيب الأولويات من البلاغة؛ تُظهره قياسات إنتاج رمز مفيد لكل وحدة مورد.

أين يمكن أن تخطئ الفكرة نفسها؟

قد يقول مهندسٌ حصيف إن نقل الدفعات إلى ساعةٍ أخرى يراكم الحمل ويرفع الذروة أو يخفض استغلال الخوادم؛ فإذا بقيت الرقائق مشتتة في الخمول لم توفر ما ظننا أننا وفرناه. وقد يجد فريقٌ آخر أن نقل حالة KV أو إعادة تحميل نموذج إلى ذاكرة رقاقة أقل ازدحامًا يستهلك كهرباء وزمنًا يتجاوزان الوفر الحراري. وقد يتغير أداء أنظمة التبريد مع درجة حرارة ماء التغذية لا حرارة الهواء وحدها. أو تكشف القياسات أن أفضل وضع للطاقة أسوأ للماء المرتبط بتوليد الكهرباء من خارج الموقع، فلا تكفي قراءة WUE المحلية.

وهناك اعتراض اقتصادي أعنف: إذا كانت Groq ونظائرها تزيد سرعة الاستدلال، فربما تزيد التطبيقات استهلاكها من الرموز بدل أن تخفض فاتورتها الإجمالية. عندما يصبح رد الوكيل أسرع، يستطيع المستخدم أن يطلب منه دورات أكثر من القراءة والتجربة والتدقيق. لذلك فإن الطاقة لكل رمز قد تنخفض بينما إجمالي الطاقة المستهلكة يرتفع، أو يتغير خليط العمل أصلاً. يبين بحث مختبر لورنس بيركلي الوطني المنشور في 2025 أن الاستهلاك المائي المنسوب إلى الأحمال يمكن أن يتباين عبر الافتراضات والمنشآت بأكثر من عشرة آلاف ضعف، مدفوعاً بكفاءة الخوادم، وكثافة الماء في كهرباء الشبكة، والاستخدام، وأنظمة التبريد والمناخ؛ ليست تلك النسبة قياساً للسعودية، بل تحذيراً علمياً من اختراع «لترات ثابتة لكل سؤال». S13

وأخيراً، إذا كان المطلوب إجابة صوتية حية لمستخدمٍ على قارة أخرى، فقد يكون القرب الشبكي أهم من أي وفر في تبريد موقع بعيد. الموقع الذي يستطيع تصدير الحوسبة عالمياً لا يصبح حتماً أفضل موقع لكل طلب عالمي. بعض التطبيقات حساس إلى تأخر الشبكة وأخرى تسمح بعمل دفعي يُرسل ويستعاد لاحقاً. أفضل صناعة لهذه الخدمة تتطلب حملين تجاريين مختلفين: فوراً يشتري أجزاء الثانية، ومرناً يشتري الطاقة الأقل تكلفة ضمن مهلة. ولا يصح تحميل النوع الأول كل مواعظ الماء التي يستطيع النوع الثاني أن يختبرها بلا إيذاء لتجربته.

ما الذي ينبغي أن يُبنى في السعودية بين 2027 و2035؟

يمكن وضع تجربة صناعية واحدة على طاولة مركز استدلال سعودي قبل إطلاق أي لقب عالمي. تُختار ثلاثة أحمال حقيقية بعقود خدمة واضحة: محادثة صوتية آتية، ومساعد برمجي متعدد الخطوات، ودفعات تقييم نموذج ذات مهلة ثماني ساعات. يُشغّل كل منها على منصة GPU وعلى منصة LPU إذا سمحت الموديلات والتراخيص والتوافر، ويُحفظ نفس معدل الجودة ونفس أعداد الرموز الصالحة ونفس توزيع طول الطلب. يُقاس TTFT عند المئين 95، وزمن الرمز بين المئينات، والعمل النافع في الساعة، وكهرباء الرف والمنشأة، وماء الموقع، وماء توليد الكهرباء إن توفرت بياناته القابلة للإسناد. ثم تكرر التجربة خلال فصول مختلفة لا خلال يوم تسويقي واحد.

لا توجد هنا دعوة لتصنيع LPU سعودي خلال ثلاث سنوات من فراغ. الإفصاح الأصلي لـ Groq كان واضحاً بشأن تطوير رقاقاتهما وتصنيعهما وتجميعهما في أمريكا الشمالية. السؤال الأعلى قيمة محلياً: هل تستطيع فرق سعودية أن تصبح متقدمة في ترجمة النماذج إلى عتاد متباين (*heterogeneous inference compilation*)، وتوزيع السياق والـ KV cache، وبناء اختبارات تجعل الطاقة والماء لكل مهمة صحيحة ضمن المهلة قابليين للتدقيق؟ هل يمكن أن تطوّر برمجيات اختيار الموقع والمنصة حسب الطول المتوقع للمخرجات وكلفة انتقال الرموز عبر الشبكة وظروف الجو، ثم تصدّر تلك البرمجيات والخبرة التشغيلية كما تصدّر خدمة الاستدلال نفسها؟ هذا مشروع بحث وتطوير ومنتج تشغيل، لا إجابة محققة حتى اليوم.

وإن أردنا نقطة انطلاق عملية أكثر تواضعاً: تكفي منظومة قياس متزامنة ومشروع A/B داخل مركز واحد يخصص نسبة صغيرة من عمله المرن لتجربة مجدول يقرأ الطقس. إن خُفّض المورد المائي دون إفساد استجابة العملاء أو نقل العبء إلى كهرباء أكثر كثافة مائية، فقد ظهر مكسب حقيقي. وإن لم يخفضه، فالنتيجة ذات قيمة؛ لأنها تقول للمهندس

أي حدّ من حدود المنظومة كان حاسماً. المهم أن يكون «النجاح» قيمة فعلية للموارد والعمل، لا حركةً جميلة في رسم بياني.

ثم تأتي الخطوة الاستراتيجية. في القرن الماضي كانت بعض الصناعات تصدّر الطاقة في صورة ألومنيوم ومعادن وأسمدة، أي تعيد تشكيل المادة بمساهمة كبيرة من الكهرباء. أما في العقود المقبلة فقد تنشأ صناعة تُحوّل الكهرباء والسيليكون والتبريد إلى زمن معرفي مدفوع الثمن: ميلي ثانية مختصرة في تفاعل صوتي، أو آلاف محاولات استدلال أتمّها وكيل علمي قبل موعد المشروع. هذا تشبيه اقتصادي مقترح لا حساباً مساوياً بين الصناعتين. في الحالتين، يصبح التفوق في تحويل الموارد إلى ناتج قابل للبيع هو جوهر المنافسة، ويصبح هدر الحرارة أو انتظار الذاكرة عيباً في السلعة لا مجرد رقم في غرفة معدات.

وهذا يعيدنا إلى المشهد الأول. كانت Groq تعرض محادثة تبدو أسرع من العادة. ولكن السؤال الاستراتيجي الذي تركته تلك المحادثة خلفها لا يخص سرعة الفهم الصناعي وحدها: ماذا لو أمكن للسعودية أن تصنع وقتاً أقل انتظاراً للعالم، وأن تعرف، على مستوى كل مهمة، الثمن الكهربائي والمائي لذلك الوقت؟ هنا يصبح مركز البيانات منشأة هندسية ذات محصول محدد، وتصبح الرقاقة أداة من أدواته، ويصبح الجدول قارئاً للطقس كما يقرأ طول الجملة. لن تجيب عن ذلك صورة رّف جديد، مهما كان لمعان المعدن حوله. ستجيب عنه تجارب تعبر من الرموز إلى الجولات المكتملة، ومن الجولات إلى العدادات، ثم تعود بنتيجة يستطيع مهندس آخر إعادة حسابها.

حدود الأدلة والمصادر

عبارات Groq و NVIDIA عن الأداء والتوسع منسوبة إلى الشركات ما لم يظهر تحقق مستقل، ونقاط الالتباس الجغرافية معلنة. جميع أرقام اختبار الـ 48 ساعة منتجة من شيفرة اصطناعية متاحة في حزمة قابلية التكرار المرافقة، وليست معلومات تشغيل حساسة ولا قياسات سعودية. القوائم الموسعة والاقتفاء لكل ادعاء في `sources.md` و `claim_ledger.csv` و `news_casebook.md`.

المراجع الرئيسية

1. [S01 Groq + Aramco Digital, 12 September 2024](#).
2. [S02 Groq, 10 February 2025, Dammam and the \\$1.5bn commitment](#).
3. [S03 Groq, HUMAIN inference provider, 27 May 2025](#).
4. [S04 Groq, The LPU Explained, 7 March 2025](#).
5. [S06 Groq and NVIDIA non-exclusive technology license, 24 December 2025](#).
6. [S07 NVIDIA, Groq 3 LPX full-production announcement, 24 August 2026](#).
7. [S08 Groq platform and network page](#).
8. [S09 Groq service tiers and batch interface](#).

.S11 US Department of Energy, cooling-water efficiency and PUE/WUE .9

S13 Lawrence Berkeley National Laboratory, water use of data-center workloads, .10
.2025

.S15 Talukder et al., Balancing Bits and Drops, arXiv 2026 .11

.S16 Li et al., PowerSlider, arXiv August 2026 .12