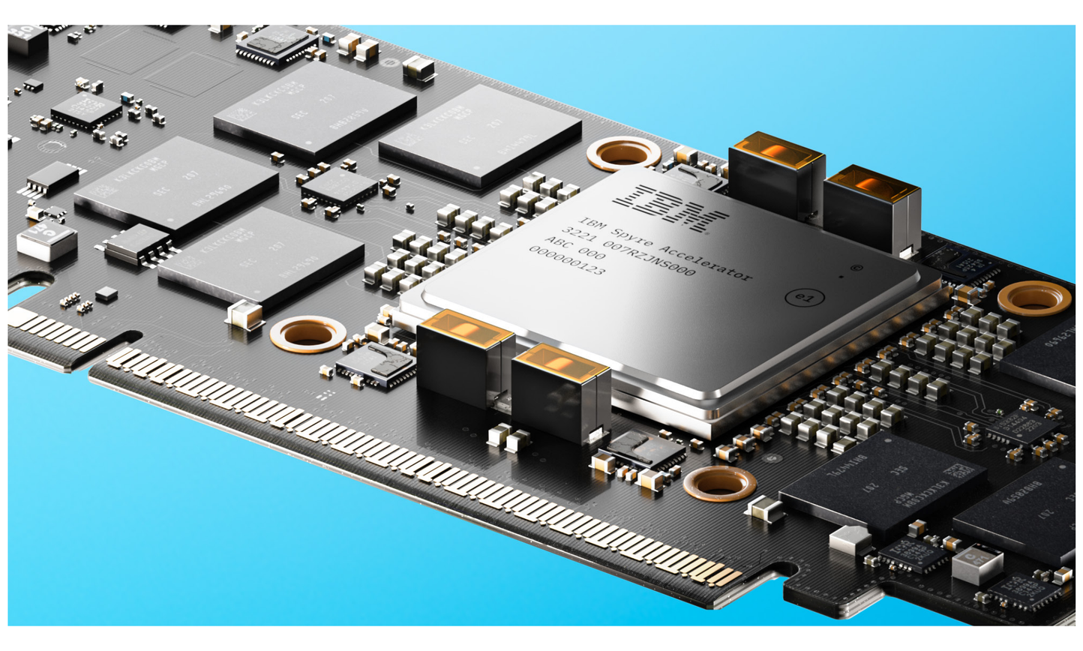
أهم 5 أسباب لتشغيل قدرات الأحمال التي يتطلبها عمل الذكاء الاصطناعي على IBM Power

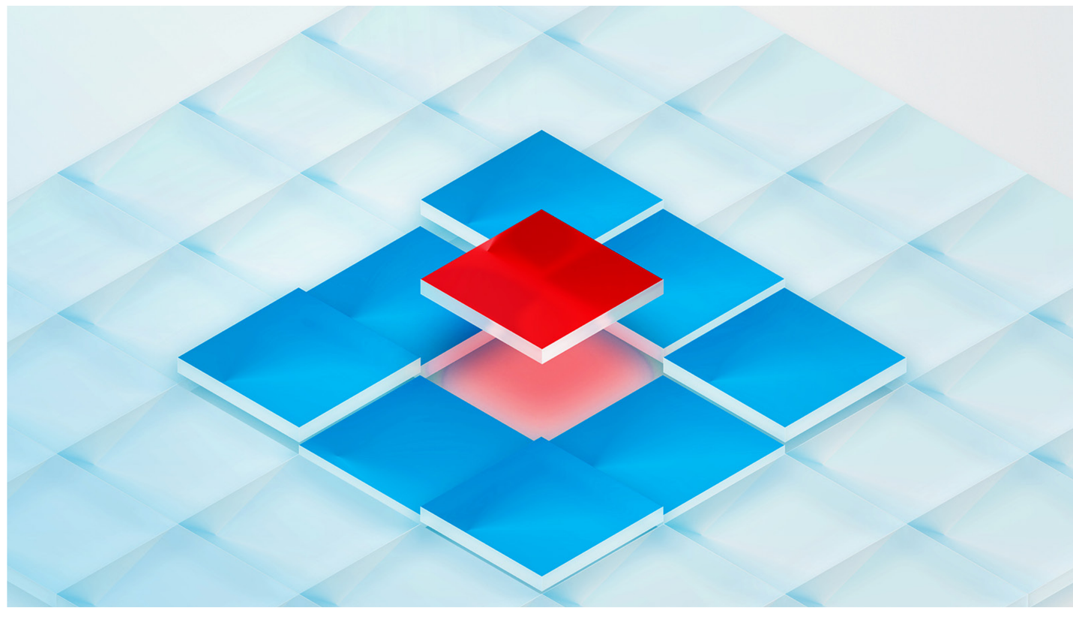
أساس موثوق به لتمكين استراتيجية الذكاء الاصطناعي

تُعد Power تحوّلًا كاملًا للأعمال بفضل التكامل السلس للذكاء الاصطناعي. سواء أكنت تعمل على دمج الذكاء الاصطناعي في المعاملات أم توسيع نماذج الذكاء التوليدي أم نشر التطبيقات عبر البيئات الهجينة، توفر لك خوادم IBM® Power الأداء والأمان والمرونة للقيام بكل ذلك دون أي انقطاع. لنلقِ نظرة فاحصة.



1 تجلب Power الذكاء الاصطناعي حيث تتواجد بياناتك

في عالم تتواجد فيه البيانات في كل مكان، في البيئات المحلية والسحابية وعلى الحافة، تحتاج إلى حلول ذكاء اصطناعي ليست ذكية فحسب، بل استراتيجية أيضًا. توفِّر Power قدرات الاستدلال بالذكاء الاصطناعي بالقرب من مكان توليد البيانات وتواجدها. لم تَعُد هناك حاجة إلى نقل البيانات الحساسة عبر الشبكات أو الاعتماد على وحدات معالجة رسومات باهظة ومعقدة. تُتيح لك Power تضمين الاستدلال بالذكاء الاصطناعي مباشرةً حيث توجد تطبيقاتك الحيوية، ما يساعد على تقليل زمن الاستجابة والمخاطر الأمنية والتكاليف الإضافية. باستخدام IBM® Spyre Accelerator وIBM® watsonx.data على Power، يمكنك الحصول على رؤى أسرع من خلال دمج بحيرة بيانات حديثة مع تسريع متقدِّم خارج الرقاقة، ما يؤدي إلى إنشاء جسر بين البيانات الموثوق بها وقدرات الاستدلال القوية.



2 توفِّر Power أمانًا على مدار الساعة طوال أيام الأسبوع

تُعَد أعباء العمل المدعومة بالذكاء الاصطناعي قيّمة، لذا لا تدعها عرضة للتهديدات. توفِّر خوادم Power حماية لأعباء عمل الذكاء الاصطناعي لديك مع أمان مدمج في كل طبقة من البنية، وتحافظ على الرؤى دون التأثير في الأداء بفضل تشفير الذاكرة الشفاف. عند التعامل مع المهام المعقدة، مثل الذكاء الاصطناعي التوليدي، يمكنك التوسع في الاستدلال بثقة دون القلق بشأن الأداء. مع خوادم Power، ستحصل أيضًا على كشف تهديدات برامج الفدية في أقل من دقيقة، وذلك بفضل IBM® Power Cyber Vault[1]، ومدة تشغيل تصل إلى 99.9999%[2] لأعباء عمل الذكاء الاصطناعي الدائمة والمرنة والمستعدة لمواجهة أي تحدٍ.

أقل من دقيقة
ضمان الكشف عن التهديدات[1]

99.9999%
نسبة موثوقية مدة التشغيل[2]

3 تمنحك Power مرونة هجينة دون أي عوائق

تُعَد المرونة الهجينة لتكنولوجيا المعلومات أمرًا حاسمًا عند نشر أعباء عمل الذكاء الاصطناعي. توفِّر خوادم Power برمجيات مؤسسية محصَّنة بالكامل لتجربة سحابة هجينة سلسة، ما يُتيح لك نقل أعباء العمل بين الأنظمة المحلية وIBM Power Virtual Server بسهولة. سواء أكنت تعمل على تدريب النماذج في السحابة أم تنفيذ الاستدلال محليًا، تمنحك Power المرونة للقيام بالمزيد. علاوة على ما سبق، تدعم هذه المنصة أكثر من 130 من أدوات وحِزم الذكاء الاصطناعي مفتوحة المصدر، مما يسهّل على الفِرق لديك بناء النماذج ونشرها وتوسيعها دون أي عوائق.

4 تقدِّم Power أداءً مستدامًا عند الطلب

من قال إن متطلبات الاستدامة لا يمكن تحقيقها مع أعباء عمل الذكاء الاصطناعي؟ تقدِّم Power11 أداءً أعلى بمرتين لكل واط مقارنةً بالخوادم المعتمدة على معالجات x86، ما يُتيح لك تشغيل أعباء العمل نفسها باستهلاك طاقة أقل.[3] وفي ظل الوضع الجديد Energy Efficient (كفاءة استهلاك الطاقة) في خوادم Power11، يمكنك تحقيق كفاءة طاقة أفضل تصل إلى 28% مقارنةً بالوضع Maximum Performance (الأداء الأقصى).[4]

أعلى بنسبة 28%
تحسين في كفاءة استهلاك الطاقة مقارنةً بخوادم x86[4]

5 توفِّر Power تسريعًا دون تقديم تنازلات

إذا كنت تتطلع إلى تشغيل الذكاء الاصطناعي على نطاق واسع، وفي الوقت الفعلي، ودون إحداث تأثير في سير العمل القائم، فإن خوادم IBM Power مصمّمة لتحقيق ذلك تمامًا، فهي توفِّر مستوى عالٍ من التوازي، وذاكرة هائلة، وتسريعًا على الرقاقة، ما يجعلها مهيأة لدمج الذكاء الاصطناعي مباشرةً في سير العمل. ولا يحتاج علماء البيانات لديك إلى إعادة صياغة التعليمات البرمجية لاستخدام المنصة؛ إذ يمكنهم تشغيل أعباء عمل الذكاء الاصطناعي كما هي. بالإضافة إلى ذلك، عند تشغيل نماذج اللغة الكبيرة، يمكنك معالجة عدد أكبر من طلبات الدُفعات يصل إلى 42% في الثانية[5] على Power S1022 -مقارنةً بالخوادم المعتمدة على معالجات x86 عند ذروة تشغيل 40 مستخدمًا متزامنًا- مع الحفاظ على زمن استدلال يقل عن ثانية واحدة.[6] وهذا يعني رؤى أسرع وعمليات أكثر سلاسة وسرعة استدلال في أقل من الثانية.

↑42%
زيادة في طلبات الدُفعات في الثانية على Power S1022 مقارنةً بخوادم x86.[5]

استمتع بتحليل عميق في مجال الذكاء الاصطناعي وتقنية IBM Power ←

1. يغطي هذا الضمان عرض التنبيه خلال أقل من دقيقة فقط. وتتم المعالجة من خلال استبدال القرص حتى تكلفة المنتج المشمول بالضمان. تنطبق الشروط والأحكام، ويمكن الاطِّلاع على التفاصيل الكاملة من هنا.
2. استنادًا إلى فترة التعطل غير المخطط لها لنظام Power E1180 واحد كما هو محسوب في أنظمة Power11 المعتمدة على المعالجات (راجع القسم: مدة تشغيل بنسبة 99.9999%) https://www.ibm.com/downloads/documents/sa-ar/10a99803d9afd776
3. . كما يستند إلى بيانات مؤشر الأداء الكمي (QPI) بتاريخ 15 مايو 2025 من IDC والمتاحة على https://www.idc.com/about/qpi واستخدام الموارد. حيث سجل IBM Power E1150 (4x30 نواة Power11 بسرعة 3.0-4.1 جيجاهرتز) QPI بقيمة 241,000E مقارنةً بخادم HPE Compute Scale-up 3200 (4x60 نواة Intel بسرعة 1.9 جيجاهرتز) QPI بقيمة 208,898، مع معدلات استخدام 75% لنظام E1150 بناءً على ضمان استغلال الأداء من IBM Power و40% لنظام x86.
– يستند استهلاك الطاقة إلى الحد الأقصى للطاقة المُدخلة: نظام IBM Power E1050 بحد أقصى للطاقة 5,200 واط https://www.redbooks.ibm.com/redpapers/pdfs/redp5684.pdf
– ، وخادم HPE Compute Scale Up 3200 بحد أقصى للطاقة 4,740 واط https://www.hpe.com/psnow/doc/a50004268enw.html?jumpid=in_pdp-psnow-qs.
4. 3. استنادًا إلى قياسات أجرتها IBM لأداء كل واط على الخوادم، حيث تم إجراء مقارنة بين وضع الأداء الأقصى ووضع كفاءة استهلاك الطاقة أثناء تشغيل أعباء عمل تعتمد على الحوسبة والأقراص والذاكرة، وذلك على أنظمة Power11، مع تكوينات مقابس وذاكرة كاملة كما يلي: E1180 بتكوين: 4 مقابس × 10 أنوية / 64 وحدة ذاكرة بسعة 64 جيجابايت DDIMM، وE1150 بتكوين: 4 مقابس × 16 نواة / 64 وحدة ذاكرة بسعة 32 جيجابايت DDIMM، وS1124 بتكوين: مقبسان (2) × 16 نواة / 32 وحدة ذاكرة بسعة 32 جيجابايت DDIMM، وS1122 بتكوين: مقبسان (2) × 16 نواة / 32 وحدة ذاكرة بسعة 32 جيجابايت DDIMM.
5. المقارنة مبنية على اختبارات داخلية أجرتها IBM لأداء الاستدلال على أسئلة وأجوبة باستخدام نموذج PrimeQA (https://github.com/primeqa، المبني على نماذج Dr. Decr وColBERT). النتائج صالحة اعتبارًا من 22 أغسطس 2023، وتم تنفيذها في ظروف مختبرية؛ وقد تختلف النتائج الفردية حسب حجم أعباء العمل، واستخدام وحدات التخزين الفرعية، وظروف أخرى. تعتمد المقارنة على إجمالي الإنتاجية بعدد الاستدلالات في الثانية على IBM Power S1022 (1x20 نواة/512 جيجابايت) مع تشغيل SMT 8 مقابل أنظمة تعتمد على Intel Xeon Platinum 8468V (1x48 نواة/512 جيجابايت). تم إجراء الاختبار باستخدام بيئات Python وAnaconda بما في ذلك حِزم Python 3.10 وPyTorch 2.0. جرى تحسين مكتبات Python المستخدمة في هذه البيئة لتعمل بكفاءة على كلتا المنصتين؛ Power وIntel. التكوين: حجم الدفعة = 60 مع 40 مستخدمًا متزامنًا. تم تحسين torch.set_num_threads(int) عبر مستويات تحميل مختلفة.
– IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html): 6.26 استدلالات في الثانية مع 40 مستخدمًا متزامنًا.
– نظام x86 المقارن: Supermicro SYS-221H-TNR (https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr): 4.4 استدلالات في الثانية مع 40 مستخدمًا متزامنًا.
– النماذج تم ضبطها بدقة بواسطة IBM على مجموعة بيانات داخلية.
6. كما استندت النتائج إلى اختبار داخلي آخر لأداء الاستدلال على أسئلة وأجوبة باستخدام نماذج PrimeQA (استنادًا إلى نماذج Dr. Decr وColBERT)، صالحة اعتبارًا من 31 أغسطس 2023، وتم تنفيذها في مختبر IBM، مع إمكانية اختلاف النتائج حسب حجم أعباء العمل واستخدام وحدات التخزين الفرعية وظروف أخرى. النتائج مأخوذة من IBM Power S1022 (2x20 نواة 2.9-4 جيجاهرتز/512 جيجابايت) باستخدام LPAR محاذي لـ NUMA للرقاقة مع 10 نوى. تم إجراء الاختبارات باستخدام بيئات Python وAnaconda بما في ذلك حِزم Python 3.10 وPyTorch 2.0، مع استخدام مكتبات Python التي تم تحسينها خصيصًا لمنصة Power. التكوين: SMT 2، torch.set_num_threads(16)؛ حجم الدفعة = 1.
– IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html)
– نماذج PrimeQA (https://github.com/primeqa).