intel.

Achieving high-impact AI without the high environmental cost

Seven tips to help you execute AI more sustainably

1 Be judicious about your AI

Evaluate potential AI projects against your business strategy and technology roadmap to focus efforts on the highest value challenges, and critically examine whether AI is necessary or if other probabilistic methods can suffice.

2 Remember: less is more

Smaller, domain-specific models can run more efficiently than generalist 'frontier' models. With fewer parameters, you can save energy from training to inference, as well as with ongoing updates.

And wherever possible, prune or compress your neural network to help reduce both compute requirements and energy consumption throughout the training and inference cycles.

3 Don't reinvent the wheel

Take advantage of repeatability and scale with pre-trained models, packaged solutions, and/or shared and federated learnings to avoid duplicating energy-intensive training.

Open APIs, like Intel® oneAPI, allow you to deploy cross-architecture solutions more efficiently, with tools, frameworks, and models helping you build once and deploy everywhere while still optimizing performance.

Studies show that many of the parameters within a trained neural network can be pruned by as much as

99%,

yielding much smaller, more sparse networks[1]

4 Consider your level of accuracy

Based on your use case, determine your tolerance for "accurate enough." With lower precision and mixed-precision techniques, rather than compute-intensive FP32 calculations, you can drive significant energy savings.

5 Optimize your hardware

By creating a more heterogeneous architecture, you can select the combination of hardware and chipsets to suit your application needs, while helping save energy across networking, storage, and compute.

Taking advantage of software optimization libraries can help ensure you're getting the best performance from your chosen hardware and applications.

Using built-in acceleration technologies can drive significant performance/watt improvements and energy savings.

Up to

10x

performance/watt improvement on AI workloads with Intel® AMX built-in acceleration vs. no acceleration[2]

6 Establish a more carbon-aware computing environment

Controlling when and where AI execution takes place can have a significant impact on the carbon intensity of your initiative, allowing you to take advantage of available renewable energy and optimize for lower grid carbon intensity.

7 Improve your cooling

Implement liquid cooling to help reduce energy consumption, as well as increase hardware lifespan, across your compute environments: on the edge or in your data center.

Liquid cooling can offer:

Up to **90%** reduction in cooling-related power use[3]

Up to **30%** increase in hardware lifespan[4]

Contact your Intel representative to learn more about how Intel can help you execute AI more sustainably

Footnotes

1. Understanding deep learning requires rethinking generalization (arxiv.org)
2. 5th Gen Intel Xeon Scalable processors using built-in Intel AMX accelerator engine deliver up to 10.2X better performance and 9.95X performance/watt improvement compared to a baseline 5th Gen Intel Xeon processor without acceleration on Image Classification with ResNet50 workloads. Performance varies by use, configuration and other factors. Results may vary . 8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x Ethernet interface, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by Intel as of 10/10/23. Software configuration: ResNet50_v1.5, Intel Model Zoo: https://github.com/IntelAI/models, gcc=11.4, OneDNN3.2, Python 3.9, Conda 4.12.0, Intel TF 2.13
3. Immersion Cooling for Data Centers | ICEraQ | GRC (grcooling.com)
4. Immersion Cooling Solutions - Lower Your OPEX and CAPEX | Hypertec