Emilio Girard

AI systems and measurement. Energy per task on robot policies, speculative decoding, inference serving on NVIDIA hardware.

Montréal, Canada · remote-first · pyloxsystems@gmail.com · emiliogirard.com · github.com/pyloxsystems · huggingface.co/pyloxsystems · English, French

I make AI models run faster and cheaper on hardware that has limits, and I measure what they actually cost in watts, joules and GPU time: robot policies on a 15 W Jetson, speculative decoding on a single GPU, serving records on a B200. Every number below traces to a run artifact, and the failed experiments are published next to the ones that worked.

SELECTED RESULTS

- **Joules Per Task** (Aug 2026): hardware energy per successful task for a vision-language-action policy on a Jetson Orin Nano, 963 J cut to 489 J with no accuracy loss and no training; energy is charged to successes, not attempts. Extended to four policies from 0.2M to 450M parameters: a 19M diffusion policy costs 5x a 30M transformer policy, and the identical vision-reuse optimization returns 18.9x on one and 1.55x on the other, so inference structure, not parameter count, sets edge energy. Code, harness and raw logs public.
- **Per-block drafter selection** (Aug 2026): 7,309 speculative-decoding blocks, four drafters verified on identical prefixes across two GPUs. Workload-level routing is a measured null (0.6% headroom); per-block confidence selection captures 61% of a 13.3% oracle gap at 43% of the cost, with bootstrap CIs, nested leave-one-workload-out threshold selection, and a stated wall-clock break-even condition. Public dataset and code.
- **Quatroo** (May 2026): 43,509 to 69,461 tok/s aggregate serving Llama-3.1-8B on a single B200, verified across two datacenters; 3,492 to 5,025 tok/s sustained single-stream. On diverse chat workloads the same stack manages only 2.21x, reported alongside.
- **Hochelaga-8B** (May 2026): first open-weights Québec French model. 85.44% accuracy on QFrCoLA against 83.85% for Claude Opus 4.7 on the same 7,546-item test split, the Opus baseline measured by me through the API. Full-parameter continued pretraining of a 24B sibling on 4x B200 under FSDP2.
- **Maxim** (Jul 2026, live at maximquebec.com): Québécois voice assistant served from a DGX Spark, 0.5 s median voice latency. Separately, a full-duplex 7B speech model answering a real Montréal phone number with mid-conversation answer injection from a larger model.
- **Connectome reservoirs** (Jun 2026): real neural wiring beats degree-matched random graphs as a reservoir computer, R^2 0.759 ± 0.014 vs 0.612 ± 0.047, 10 of 10 seeds; two null results across 14 connectomes reported with it.
- **Jetson reproductions** (May 2026): three published papers ported to TensorRT INT8/FP16 on a Jetson Orin Nano with added energy measurements; five-seed class-incremental ablation; 263 fps terrain navigation.

TIMELINE

Sep 2026	Starting a college AI program in Montréal alongside the lab work.
Aug 2026	Joules Per Task and per-block drafter selection: two papers with public code and data.
Jun to Jul 2026	Full-duplex phone line; connectome experiments; Maxim goes live on the public web.

May 2026	Jetson paper reproductions; Hochelaga-8B; Quatroo B200 serving records.
Feb to Apr 2026	Eva, a WhatsApp coordination agent, in daily production at a construction firm. Pylox Vision, an 8-camera Jetson detection and re-identification stack, at a pilot site (74K events). First public fine-tunes on Hugging Face, each benchmarked against its base; three lose and say so.
Nov 2025	Pylox Systems Inc. starts as a one-person lab around a DGX Spark (GB10).
Before	DEP in Machining Techniques, Rosemount Technology Centre, Montréal. Operator on a Mazak Integrex e-670H at Mecaer America (Laval), machining F-35 landing gear steering head components in titanium.

HARDWARE

Jetson Orin Nano (x2)	energy measurements; TensorRT INT8/FP16 ports; MAXN tuning, Tegra cuBLAS quirks
DGX Spark (GB10, sm_121)	daily driver: 24B full-parameter CPT, EAGLE-3/DFlash drafter training, NVFP4 serving, live voice product
RTX 5090, RTX Pro 6000	33.8x warm sustained speculation; counterfactual collection; production voice serving; 30B diffusion-LM experiments
A100, H100, B200 (up to 4x)	full-duplex speech serving and training (FSDP, two cards); 24B pretrain; Llama-8B serving records
Raspberry Pi 5, Apple M2/M4	stripped-OS daemons; Metal targets; SwiftUI clients; 42-CPU Ray cluster across the fleet

STACK

Inference	TensorRT, TensorRT-LLM, vLLM, SGLang, llama.cpp, ONNX Runtime; NVFP4, MXFP4, FP8 KV-cache, INT8 QDQ with calibration, GGUF
Speculative decoding	EAGLE-3, DFlash, Medusa, MTP, suffix automaton, n-gram, grammar-constrained trees, self-distilled drafters, SpecForge
Training	full-parameter CPT, SFT, DPO, GRPO, distillation, FSDP2/DTensor, LoRA/QLoRA/DoRA/PiSSA/rsLoRA, NeMo, Unsloth, LLaMA-Factory, Axolotl
Speech and vision	Moshi/Mimi full-duplex, PersonaPlex, F5-TTS, Whisper, NVIDIA Canary; YOLO v8 to 11, RT-DETR, OSNet Re-ID, Frigate, COLMAP, gaussian splatting, RoomPlan LiDAR
Systems	CUDA, PyTorch, Ray, ROS 2, Docker, systemd, Ansible, Cloudflare, Tailscale, Telnyx telephony, MQTT, Qdrant, Neo4j, Postgres, Node.js, FastAPI, SwiftUI

PUBLIC ARTIFACTS

github.com/pyloxsystems: joules-per-task, per-block-drafter-selection, risk-aware-terrain-jetson, cbcl-pr-jetson, pylox-vision-demo. huggingface.co/pyloxsystems: 10 models including hochelaga-8b, hochelaga-8b-nvfp4 and hochelaga-8b-eagle3-drafter. Full ledger with the failures: emiliogirard.com/plain.html.