TECHNICAL MEMO · 16 NOVEMBER 2026

Doing More With Less: Symbiotic Compute Beyond Computer Vision

Modern AI infrastructure defaults to brute force: larger models, hotter GPUs, liquid cooling, and multi-year queues for grid power. In orbit—and across power-constrained edge environments on Earth—that playbook breaks down because every watt of power and heat must be handled by the structure itself.

At Project Vela, we are developing solar-integrated orbital compute around a simple physical principle: moving information is far cheaper than moving heat. Step 1 on our roadmap (V1 Pathfinder) measures how much real-world capability can be unlocked from commercial low-power accelerators before introducing our custom V2 stream-controller chip.

To test those limits, we built a 40-chip desktop bench and Vela Node 1—a compact compute node pairing a six-core Arm Cortex-A55 host with three four-chip commercial edge AI accelerator M.2 modules (12 chips total). We are not building a vertical weather vendor or an OCR shop. Our goal is architectural: proving that weight-stationary chips designed for computer vision can run attention, language agents, and spatiotemporal physics at a fraction of standard energy when co-designed with an active host.

1. Hardware Foundation: Bit-Exact Scaling and Dynamic Power

Across commercial edge AI accelerator arrays and our C++ streaming runtime, deterministic accuracy and power scale cleanly with workload demand.

Hardware foundation measurements

Workload or Test	Hardware	Throughput	Accuracy vs. FP32	Power/Energy
10-min full-rate soak	40 accelerator chips	8,482 fps	5,089,557 bit-exact	Peak 66 °C
ResNet-50	4 to 40 chips	692 to 6,924 fps	100.2% of FP32	10.8 mJ/frame
LLM layer	4 accelerator chips	343 fps	Within 1% ppl	6.1 mJ/frame
Vision & EO suite	4 accelerator chips	21 to 692 fps	7 of 9 builds match	0.35–1.9 W/chip

Power above idle tracks arithmetic load—from 0.36 W per chip on small MLPs to 1.84 W per chip on ResNet-50. Across overhead ship detection (0.955 mAP50 on ships) and 16-bit pre-downlink cloud screening (93.2% vs. 94.1% FP32), the chips match full-precision baselines.

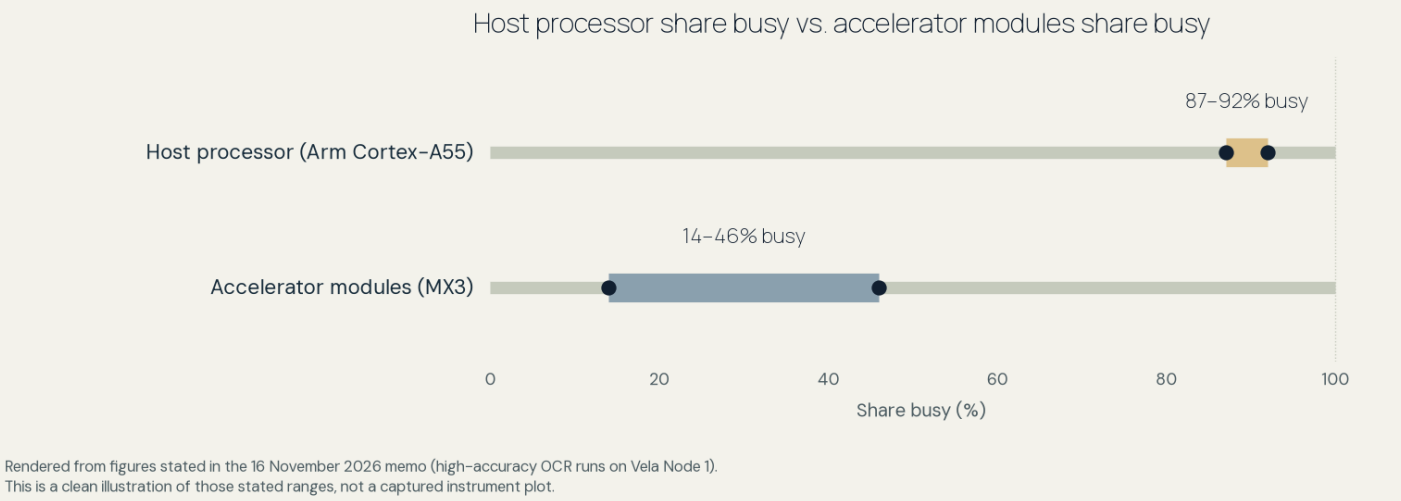
> The deck visual “Seven of nine builds score the same on the accelerator chips as at full precision” was not in the source charts folder, and the memo does not list the nine individual scores, so no substitute chart is shown here. The table above retains the stated result: 7 of 9 builds match.

2. Attention, Document Intelligence, and Multi-Chip Language Agents

Weight-stationary chips multiply live data by fixed weights rather than multiplying two live streams as attention requires. Reformulating attention into native accelerator operations unlocked both document pipelines and distributed language models:

- **Attention OCR on Node 1:** Moving an attention-based text reader from Arm cores to Node 1’s 12 accelerator chips raised steady receipt throughput from 745 to 10,345 pages/hour.
- **End-to-end field accuracy:** Adding an upgraded on-chip detector and field picker reached 77.4% accuracy at 9,679 pages/hour, or 80.5% at 6,872 pages/hour with host spell-checking.

- **Host bottleneck exposed:** During high-accuracy OCR runs, Arm cores ran 87% to 92% busy while accelerator modules sat 14% to 46% busy waiting on the host—directly motivating our V2 controller chip.
- **40-chip language agents:** Splitting all 24 layers of Qwen2.5-0.5B-Instruct across 40 chips delivered 1,607 tokens/second across 2,048 sessions and finished 1,800 agent tasks within 0.89 points of FP32 (0.33 with retry).



Rendered from figures stated in the 16 November 2026 memo (high-accuracy OCR runs on Vela Node 1).
This is a clean illustration of those stated ranges, not a captured instrument plot.

Host processor share busy vs. accelerator modules share busy. Rendered from figures stated in this memo (high-accuracy OCR runs on Vela Node 1). This is a clean illustration of those stated ranges, not a captured instrument plot. A matching deck visual was not in the source charts folder.

3. Case Study: Symbiotic Storm Triage and Nowcasting

Satellite and radar sensors generate far more pixels than edge power or downlink budgets can brute-force. Using the full public SEVIR benchmark (~35,700 training sequences; 12,159 test sequences) and whole-U.S. NOAA/SPC archives, we co-designed a two-part weather pipeline across Node 1's Arm host and accelerator chips.

Part A: 1-Hour Radar Nowcasting at ~27× Lower Energy

We trained a 16.6M-weight convolutional U-Net (65 minutes of radar in, 60 minutes out across 12 frames) and fused a validation-fitted [12, 1024] intensity calibration lookup directly into an Arm NEON output decoder on Node 1.

Nowcasting skill, speed, and energy

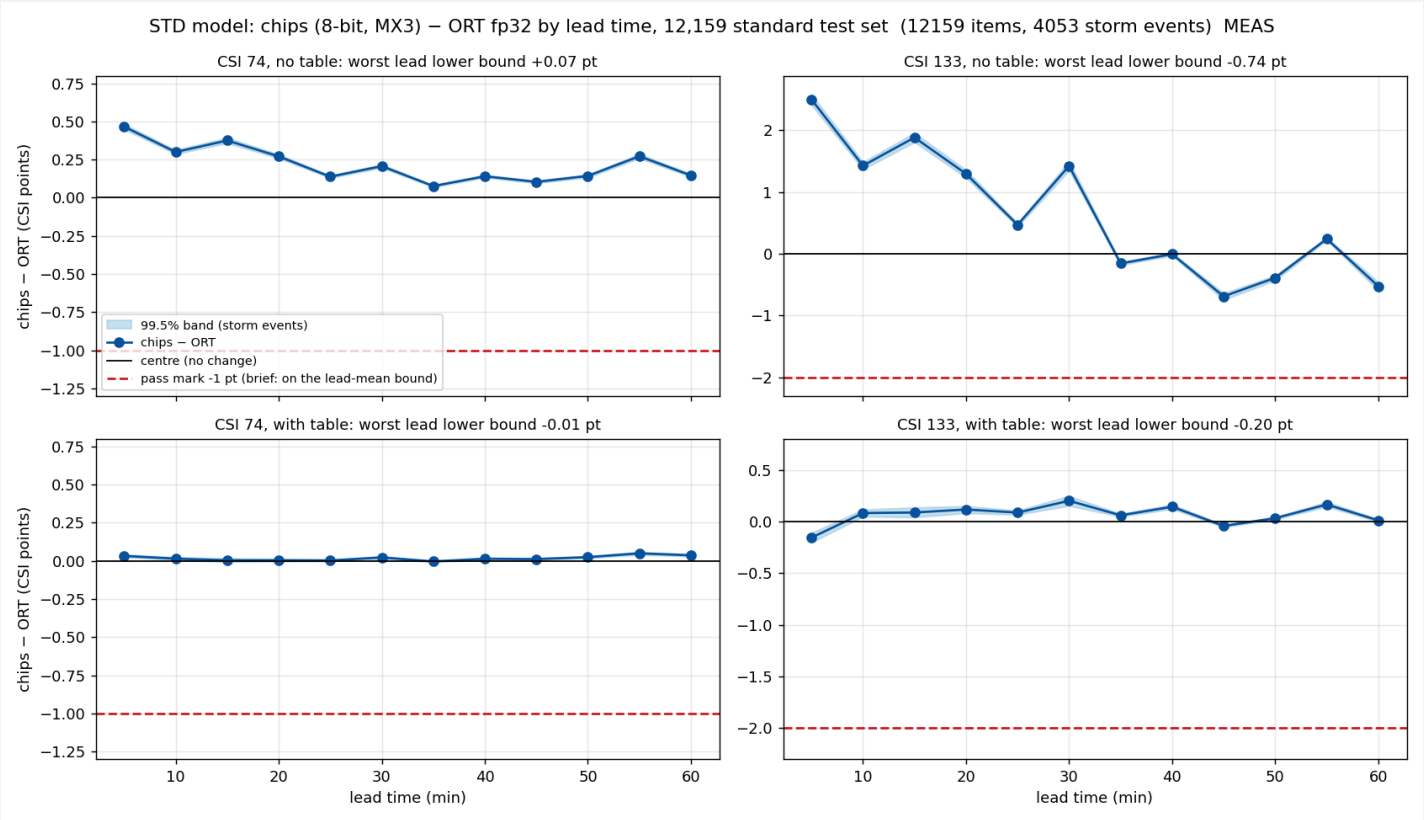
System & Precision	Hardware	CSI-M (6 levels)	Speed	Energy/Forecast
"Nothing changes"	Baseline	0.262	—	—
Earthformer (as published)	RTX 5080	0.431 (0.442 paper)	30 fps	9.3 J (card only)
Earthformer + host remap	RTX 5080	0.449	30 fps	9.3 J (card only)
Vela U-Net + remap (FP32 / FP16)	RTX 5080	0.441	245 / 546 fps	0.86 / 0.50 J
Vela U-Net + NEON remap (8-bit)	Vela Node 1	= bench chips (byte-identical, first 2,000)	46.8 forecasts/s	~0.34 J (single measurement)

On the bench accelerator chips, the same standard model over the full 12,159-sequence test, with the calibration table, stayed within +0.53% pixel error and +0.07 heavy-rain points of full precision. Six-level skill at 16 km was 0.482 on the chips versus 0.475 at full precision.[a] Node 1 skill on the first 2,000 sequences equals those bench-chip forecasts (byte-identical).

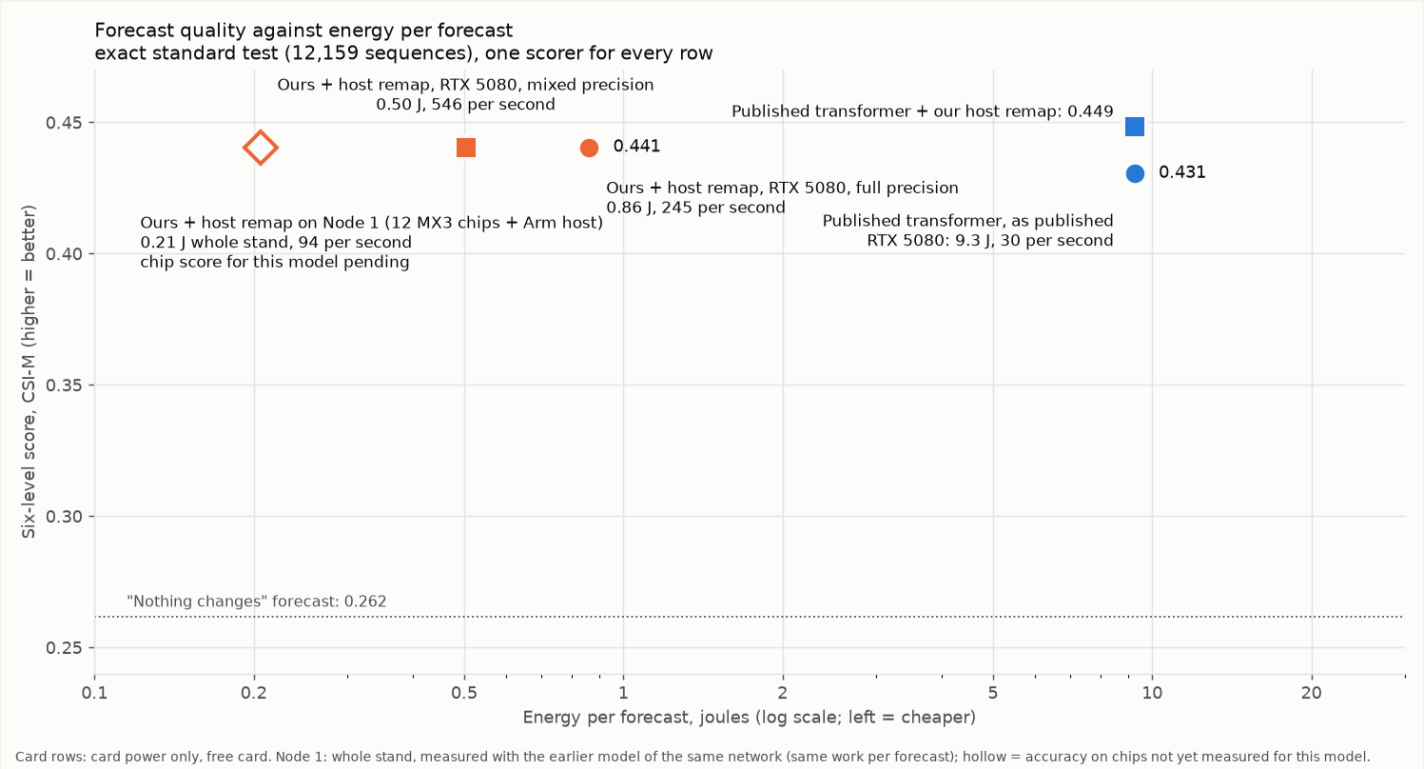
[a] Six-level skill at 16 km blocks, chips vs ORT fp32, with the calibration table, measured on the 12,159-sequence standard test.

- **Transformer-class skill at ~1/8th the math:** Performing 33 GFLOPS vs. 257 GFLOPS, our co-designed model beat the hosted Earthformer checkpoint as published (0.441 vs. 0.431; 0.457 vs. 0.431 at 4 km blocks) and reached 98% of Earthformer even when gifting it our host remap.

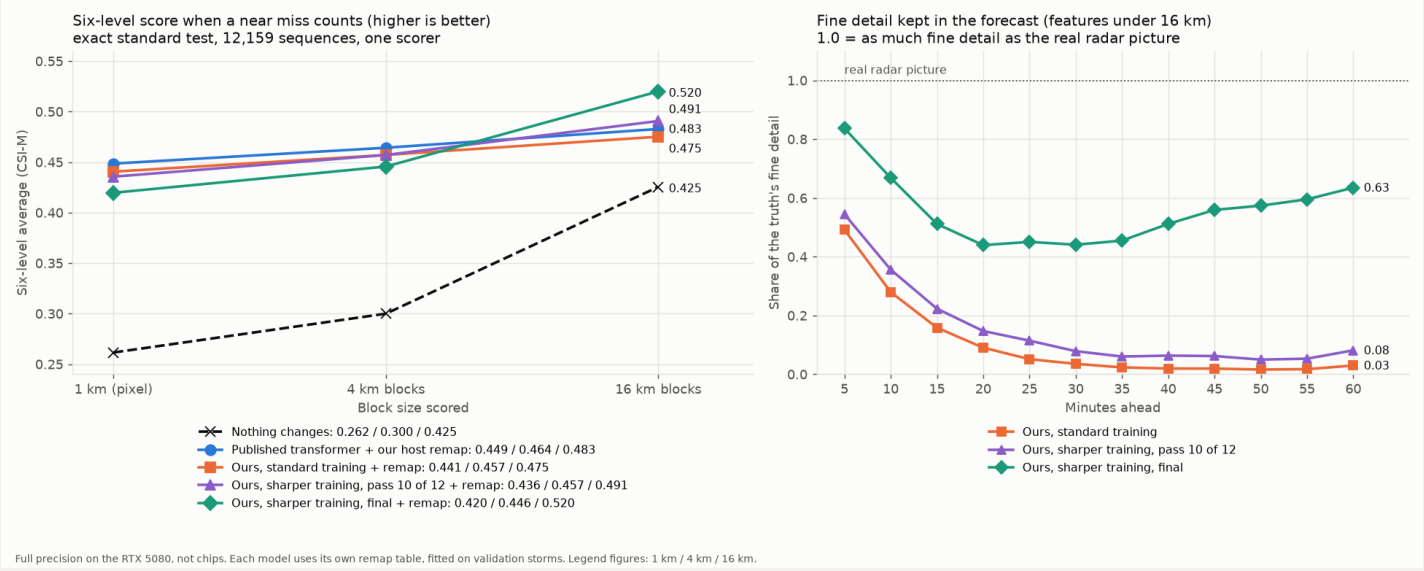
- **Bit-exact NEON fusion on Node 1:** Vectorizing output decode with fused remapping on Cortex-A55 added just 7.0 ms of host work (vs. 20.1 ms unfused). The first 1,000-sequence chunk on Node 1 ran at 46.8 forecasts/s and ~0.34 J per forecast on the metered stand (a single measurement; ~27× below Earthformer's 9.3 J GPU-only energy). The first 2,000 sequences matched the bench chips byte for byte.
- **Standard model on the bench chips:** Across the full 12,159-sequence test, with the calibration table, the 8-bit accelerator run stayed within +0.53% pixel error and +0.07 heavy-rain points of full precision. Six-level skill at 16 km was 0.482 on the chips versus 0.475 at full precision.
- **21× more fine storm detail (Fidelity Variant):** Training the same network with a Fourier amplitude/correlation target preserved 63% of sub-16 km storm detail at 60 minutes (vs. 3% for standard training), beating the remapped transformer on 16 km blocks (0.520 vs. 0.483) while trading 2.1 points at strict 1 km pixel overlap. On a 1,872-sequence chip check, the same fidelity model passed all five marks with the calibration table: the worst heavy-rain bound was +0.23 points versus full precision, and all 60-minute fine detail was kept. Without the table, heavy rain dropped 1.4 points.



Chips minus ORT fp32 by lead time on the 12,159-sequence standard test (4,053 storm events). Source chart: std_chart1_spc_by_lead_12159.png (MEAS). With the calibration table, the heavy-rain (CSI 133) series stays close to full precision across lead times; the worst lead lower bound is -0.20 points.



Forecast quality against energy per forecast on the exact standard test (12,159 sequences). Source chart: benchmark_energy.png. Caption correction: this figure still plots an earlier Node 1 point at ~0.21 J and ~94 forecasts/s, with skill marked pending (hollow marker). Those figures are superseded: the measured first-chunk rate is 46.8 forecasts/s at ~0.34 J (single measurement, ~27× below Earthformer's 9.3 J), and Node 1 skill on the first 2,000 sequences equals the bench chips (byte-identical). The image itself was not redrawn.



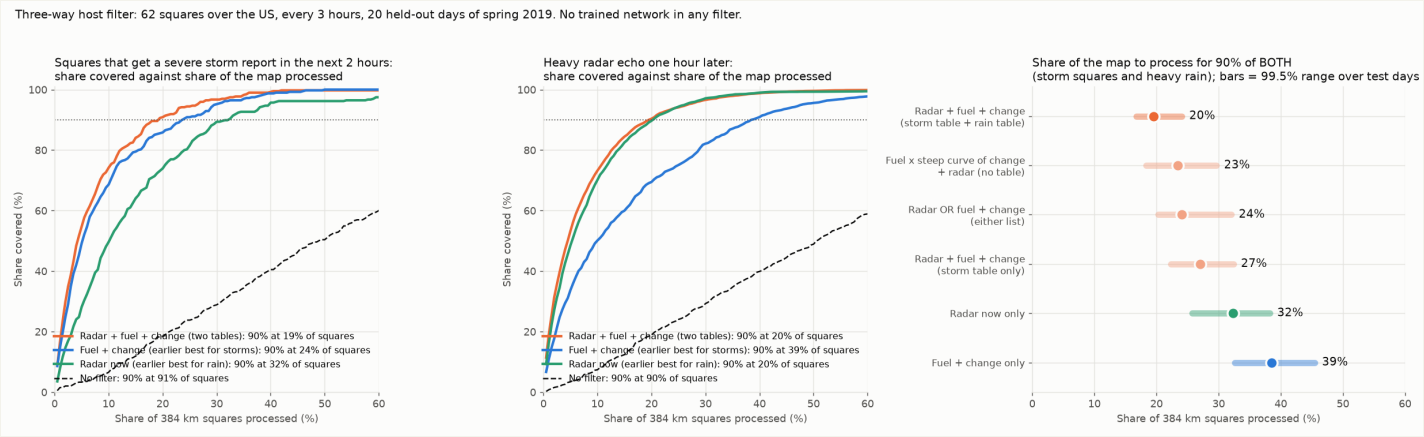
Skill by block size and fine storm detail retained. Source chart: sharper1_std.png. Full precision on the RTX 5080, not chips, as noted on the figure.

Part B: Zero-Network Host Triage—70% Less National Cycle Energy

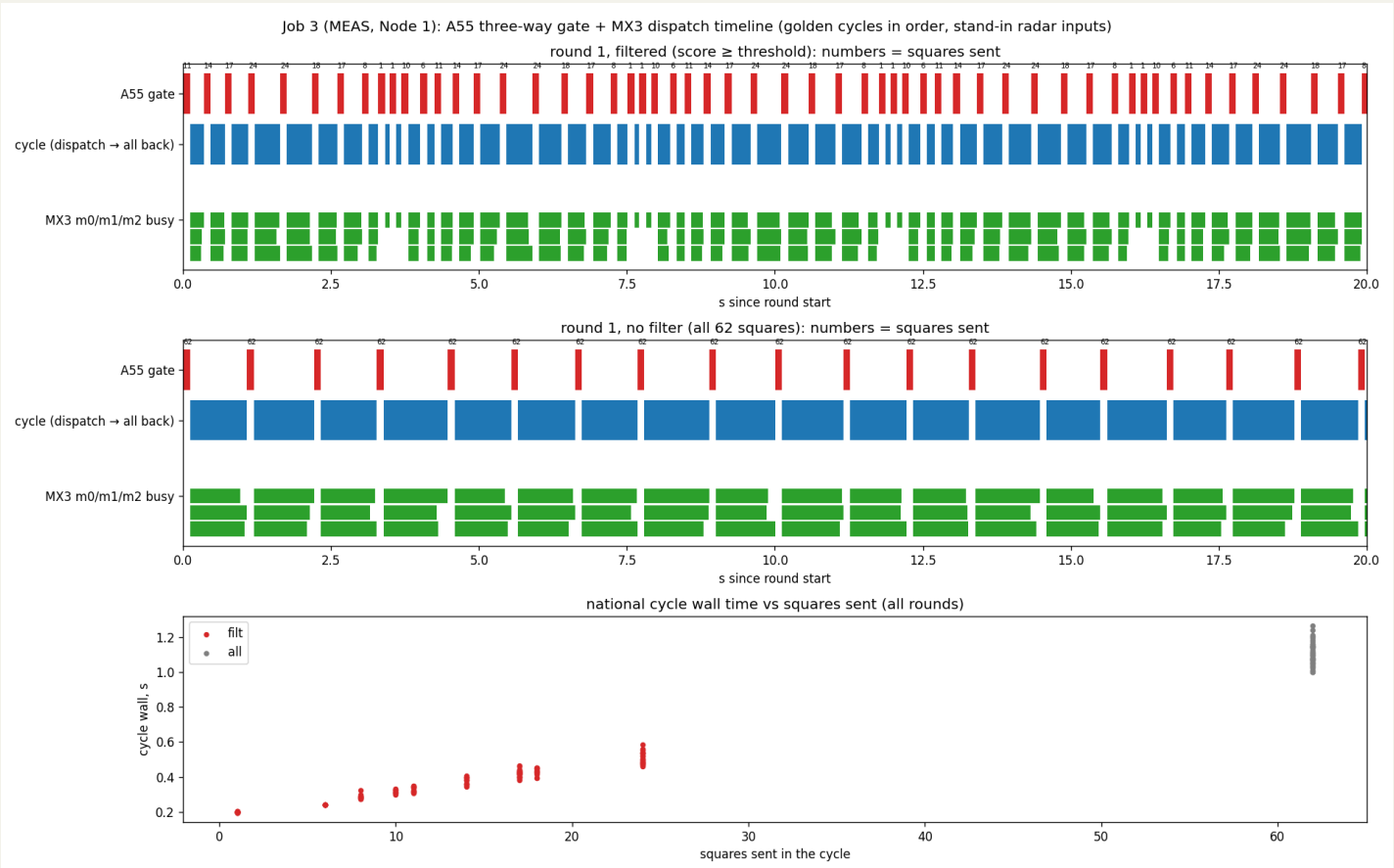
Instead of running neural networks blindly over calm sky, Node 1's Arm host runs a zero-network physics pre-gate combining current radar echo, storm fuel, and two-frame spatial boundary shifts (requiring zero multi-day baseline history):

- **11× storm breakout separation:** Among cells with top-5% storm fuel across 1.9 million Spring 2019 cell-times, fast-changing boundary cells were 11× more likely to spawn severe weather within two hours than calm cells with equal fuel.
- **80% of calm sky discarded:** Across 62 U.S.-wide 384 km squares on 20 held-out days, the three-way host gate captured 90% of both severe storm breakouts (0–2h) and heavy rain (1h) while waking accelerators on just 20% to 23% of the map.

- **Measured on Node 1 silicon:** Evaluating all 62 U.S. squares took 114 ms on one Arm A55 core and dispatched ~13 active squares per pass to the 12 accelerator chips—cutting national-pass wall time from 1.10 s to 0.36 s and metered stand energy from 18.21 J to 5.45 J (-70%).



Three-way host filter on 62 U.S. squares, every 3 hours, 20 held-out days of Spring 2019. No trained network in the filter. Source chart: host_filter3.png.



Gate and dispatch timeline per national cycle on Node 1 (stand-in radar inputs). The memo cites gate_and_dispatch_timeline_per_national_cycle_2.png; that filename was not in the charts folder. This is the closest file, gate_and_dispatch_timeline_per_national_cycle.png.

4. Measurement Boundaries and Next Steps

We report hardware measurements with explicit boundaries:

- **Power metering scope:** RTX 5080 figures reflect GPU board power only (excluding host PC draw, favoring the GPU), while Node 1 stand figures meter the 12 V carrier stand; inline Arm host + third-slot supply metering is part of Stage 4 verification.
- **Seasonal & live validation:** Triage filters were scored on 20 held-out Spring 2019 days using hourly fuel grids and stand-in radar arrays for Node 1 dispatch timing; multi-season validation and live radar feeds are queued next.
- **Confidentiality:** Benchmarks ran on commercial edge AI accelerator silicon. Attention graph layouts, weight preparation, and multi-chip orchestration remain proprietary to Alien Alliances, Inc. and are shared under mutual NDA.
- **Node 1 nowcast run:** The first 2,000 sequences completed and matched the bench chips byte for byte. The run then stopped on the second 1,000-sequence chunk after an accelerator write timeout (**wait timeout, retrying**). PCIe flags, temperatures, and post-run goldens were clean. The best unproven guess is that the SD card was busy writing the next 1.9 GB input chunk. The 46.8 forecasts/s and ~0.34 J figures are a single measurement from the first chunk, which overlapped that write, and are not a completed multi-round rate.

Inquiries and NDA data-room requests: info@project-vela.com.