INSTITUTE FOR PHYSICAL AI @ BMI
The Charlot Lab & The Hiner Lab

Technical Report TR–2026–04
Survey / Review · Preprint v1
6 July 2026

ONE FLUID, THREE CURRENTS

Microfluidics as a Shared Substrate for Physical AI: Co-Optimizing Heat, Charge, and Information in One Electrolyte Network — A Survey and Research Position

The seam where computation meets its energy budget: the same embedded fluid that must remove heat and deliver power can also carry a slow, local layer of computation.

David Jean Charlot, PhD · Michael Hiner

The Charlot Lab & The Hiner Lab, Institute for Physical AI @ BMI

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org
Interactive companion: physicalai-bmi.org/research/hiner-lab#topic-microfluidic · Formal program: Defensive Publication DP–2026–01

Abstract. An embodied system carries its power plant with it. At the edge, the dominant constraints on computation are not clock speed but the two currents that keep a chip alive: the heat it must reject and the power it must be fed. This report surveys a body of work in which those two currents already flow through a fluid — embedded microchannel coolant, and on-chip electrochemical flow cells — and observes that the same electrolyte can carry a third: information, through iontronic nanofluidic devices whose conductance depends on ion history. The report organizes the field into three currents through one network, reviews the maturity of each, and states a research position: because the three share one flow and one dissipation budget, they should be priced jointly, in exergy rather than energy, and any design that optimizes cooling, power, or computation in isolation is dominated whenever the physical couplings between them are nonzero. We are explicit about maturity. Embedded fluidic cooling is an engineered technology; on-chip fluidic power is a demonstrated laboratory capability; in-fluid iontronic computation is early-stage research whose devices operate at millisecond-to-second timescales and are therefore unsuited to replacing a processor. The honest opportunity is not a faster computer but a slow homeostatic layer co-located, at near-zero marginal exergy, in the thermal and power skin an embodied machine must already have. Section 2 states the method; Sections 4–6 survey the three currents; Sections 7–9 give the couplings, the exergy objective, and a determinism constraint that keeps the readout bit-exact; Section 10 is a maturity assessment. The

full optimization program is released separately as a public-domain defensive publication. This report reports no new experimental measurements.

1. Introduction

A datacenter can treat cooling and power as facilities problems, solved at the building scale and largely out of the designer's view. An embodied machine cannot. A legged robot, a surgical tool, a wearable, or an implant carries its energy supply and its heat sink inside the same volume as its computer, and every watt spent moving heat or conditioning power is a watt not spent moving the body. For Physical AI, the design metric is not operations per second but useful work per joule delivered into a body's budget, and at that metric the thermal and electrical infrastructure is not a facilities detail but a first-class part of the machine.

Two independent engineering communities have converged on the same medium to carry that infrastructure. To remove heat at the densities modern silicon reaches, coolant has been moved off the heat spreader and into microchannels etched into the chip itself, co-designed with the circuitry rather than bolted on afterward[1]. To deliver and regulate power without bulky conversion, electrochemical flow cells have been integrated on-chip, in some designs sharing the very coolant loop that removes the heat[2,3]. In both, a fluid — typically an aqueous electrolyte — already threads through the active region of the device.

A third community, working on nanofluidic iontronics, has shown that an electrolyte in a confined channel is not only a coolant and a conductor but a computational medium: channels whose conductance depends on the recent history of the ions inside them behave as memristors, the history-dependent circuit element first posited in 1971[4], now realized in water[5] and assembled into networks that perform brain-inspired temporal computation[6,7]. The three lines of work have proceeded largely apart. This report reads them together, because on an embodied device they are not separate: they can be the same fluid.

2. Scope and method

This is a survey with a stated research position, not an experimental paper. It reviews published, peer-reviewed and preprint work across three fields — embedded microfluidic cooling, on-chip electrochemical power, and nanofluidic iontronics — and the classical transport and thermodynamics that connect them[8,9]. For each claim it favors a primary source, and it is explicit about the maturity of each capability, because the three differ by many years of development. The report advances a position in Sections 7–9 and Section 11; it reports no original measurements. The interactive companion referenced throughout is an illustrative first-order model, not a validated simulator, and is labeled as such on the site. The formal statement of the joint optimization program — decision variables, governing equations, objective, and constraints — is released as a separate public-domain defensive publication (DP-2026-01) so that it stands as prior art; this report is the readable survey and argument around it. Its principal limitation is that the third current, in-fluid computation, rests on a

young and fast-moving device literature, and several quantitative figures for it are best-estimates rather than settled values; these are marked where they appear.

3. Three currents, one network

The organizing picture is a single embedded electrolyte network that carries three currents at once (Figure 1). **Heat flows out:** the fluid is a coolant, convecting away the power the chip dissipates. **Charge flows in:** the same fluid is the working medium of an electrochemical flow cell, delivering or locally regulating power. **Information flows through:** where the channel narrows to the nanoscale, the electrolyte's ion configuration is a memory, and reading its conductance is a computation. Table 1 states the transfer principle, the rough magnitude, and the maturity of each. The three do not merely coexist; they are coupled through shared physics, developed in Section 7. That coupling is the reason the report argues for a joint objective rather than three separate ones.

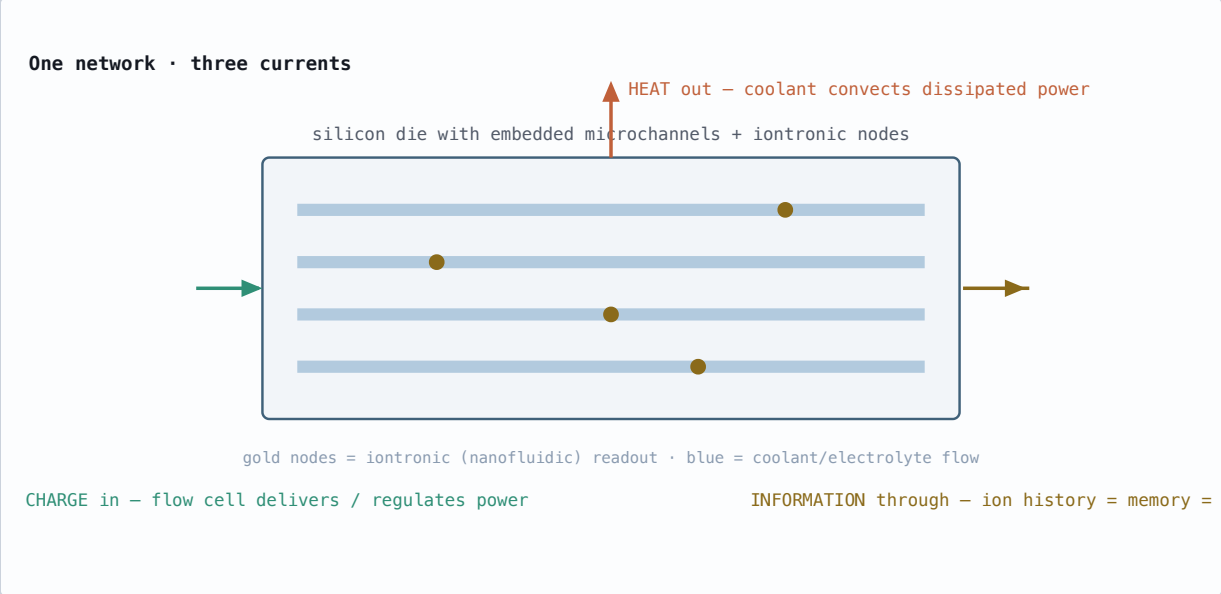

Figure 1. The organizing picture. One embedded electrolyte network is coolant, power medium, and computational substrate at once. Heat leaves with the flow; charge is delivered by electrochemistry in the same fluid; information is held and read in the ion configuration at nanoscale nodes. An interactive version that lets the reader trade flow rate against temperature, power, and readout determinism is at physicalai-bmi.org/research/hiner-lab.

Table 1. The three currents through one network: principle, rough magnitude, and maturity. Magnitudes are order-of-magnitude and drawn from the cited sources; the compute figures are best-estimates from a young literature.

CURRENT	PHYSICAL PRINCIPLE	ROUGH MAGNITUDE	MATURITY
Heat out	Forced convection in embedded microchannels; co-designed with the circuit	heat flux > 1.7 kW/cm² at low pumping power[1]	Engineered technology
Charge in	Membraneless co-laminar / redox flow cell in the coolant loop	power density order 0.1–1 W/cm²[3]	Lab-demonstrated

Information through	Nanofluidic iontronic memristor; conductance depends on ion history	< 1 V drive; memory over ms–s[5]	Early research

4. Heat out: embedded microfluidic cooling

As on-chip power density rose, the thermal path became the limit, and the response was to shorten it. Rather than conduct heat up through the package to a remote cold plate, coolant is brought into microchannels etched directly into or immediately beneath the active silicon, so the fluid meets the heat where it is generated. The decisive step for this report is co-design: van Erp and colleagues fabricated the microfluidic cooling and the electronics in the same substrate, matching the channel network to the map of where heat is actually produced, and demonstrated the removal of heat fluxes above 1.7 kW/cm² while spending very little pumping power[1]. Subsequent work has pushed manifold and channel geometries for large-area, high-power dies[10]. The lesson that carries forward is that the coolant network is a designed part of the chip, its geometry a free variable, and its flow already present in the active region — precisely the conditions under which a second or third use of the same fluid costs little to add.

5. Charge in: fluidic power and regulation

Power delivery is the mirror image of heat removal, and it too has been moved into a fluid. Membraneless microfluidic fuel cells and redox flow cells exploit co-laminar flow — two streams that flow side by side without turbulent mixing — to run an electrochemical half-reaction at each electrode and produce power on-chip without a separator membrane[3]. Reported power densities are of order 0.1–1 W/cm², modest against a processor's demand but useful for local regulation and for offsetting delivery losses. The development that matters here is integration with cooling: recent work operates the coolant loop and a redox flow cell as one system, using the backside microchannel network simultaneously to cool the chip and to carry the electrolyte that powers it, so a single fluid loop does thermal and electrical duty at once[2]. This is the first coupling made concrete in hardware: the fluid that removes heat is the fluid that delivers charge, and its flow rate sets both.

6. Information through: nanofluidic iontronics

The third current is the least mature and the most interesting. When an electrolyte-filled channel is narrowed to the nanoscale, ion transport through it is governed by the coupled Poisson–Nernst–Planck equations, and in appropriately shaped channels the channel's conductance comes to depend on the recent history of the voltage applied across it. That history-dependence is exactly the defining property of the memristor[4], and Kamsma and colleagues demonstrated it in a conical microfluidic channel, an aqueous iontronic memristor whose current–voltage curve is a pinched hysteresis loop and whose state persists over a physically meaningful time[5]. Networks of such devices perform temporal computation: because a volatile iontronic memristor is naturally a leaky integrator, a fixed random network

of them implements a physical reservoir computer, with only a linear readout trained, and the group has demonstrated brain-inspired signal processing and an explicit mapping to echo-state and band-pass networks driven directly by physical inputs such as pressure[6,7].

Two properties make this a candidate for the embodied edge rather than the datacenter. First, the devices operate at low voltage and their dynamics are set by ion relaxation, so switching energies are small — plausibly sub-picojoule per event, though this is a best-estimate from a young literature and depends strongly on geometry. Second, and decisively, their natural timescale is milliseconds to seconds, matching the timescale of sensing, homeostasis, and reflex rather than of digital logic. That is the same property read two ways: a limitation if one wants a fast computer, and a match if one wants a slow, physical, sensor-adjacent layer that computes on the timescale the body actually operates on.

7. The couplings

The three currents are not independent, because they share the fluid, its flow, and the temperature field. Three couplings are worth naming; all follow from classical transport and electrochemistry, not from any proprietary mechanism[8].

K1 — heat helps charge (recovery). Electrochemical kinetics are temperature-activated: the exchange current density of an electrode reaction rises with temperature through an Arrhenius factor, while the equilibrium potential shifts with temperature through the Nernst relation. In compact form the delivered power responds positively to temperature,

$$\frac{\partial P_{\mathrm{fc}}}{\partial T} > 0 \quad \text{via} \quad j_0(T) = j_0^{\mathrm{ref}}\, e^{-\frac{E_a}{R}\left(\frac{1}{T}-\frac{1}{T_{\mathrm{ref}}}\right)},$$

so the heat that must be removed can, within limits, upgrade the charge that is delivered: dissipation becomes partly a credit rather than a pure loss. This is why the report prices in exergy, which accounts for the quality of recovered heat, rather than in raw energy.

K2 — one flow, three opposed demands (adversarial). A single scalar, the flow rate Q, sets cooling, power, and pumping cost at once, and it cannot satisfy all three. Faster flow cools better and reduces the residence time that limits electrochemical utilization, while pumping power grows steeply:

$$\frac{\partial T_{\max}}{\partial Q} < 0, \qquad \frac{\partial \xi}{\partial Q} < 0, \qquad \frac{\partial P_{\mathrm{pump}}}{\partial Q} > 0.$$

Three partial derivatives of opposite sign in one variable is the core operating tension, and it is precisely what a joint objective must resolve and a per-current objective cannot see.

K3 — heat and information are the same field (bidirectional). Two effects tie temperature to computation. A temperature gradient drives an ion flux by thermodiffusion, the Soret effect, so the thermal-management field can itself write ionic state,

$$\mathbf{J}_s^{\text{Soret}} = -\,D_s\,c_s\,S_{T,s}\,\nabla T,$$

while at the same time the computation is a distributed heat source, since every switching event dissipates into the very fluid that is cooling the chip. Heat is therefore both an input to and an output of the computation, which is what makes cooling and computing non-separable in principle rather than merely inconvenient to separate in practice.

> Because K1–K3 are generically nonzero, the three currents cannot be correctly priced one at a time. A cooling design that ignores K1 discards recoverable exergy; a power design that ignores K2 mis-sets the one flow that also governs temperature; a compute design that ignores K3 treats as noise a thermal field that is in fact writing its state. The couplings are the argument for a single objective.

8. Pricing in exergy, not energy

Energy is conserved and therefore a poor scorecard: it does not distinguish a joule of delivered work from a joule of waste heat at ambient. The right currency is exergy — available work — which charges for the quality of each stream and credits recovered heat at its true worth[9]. The joint objective the report advocates is net exergy destroyed per relevant outcome: sum the compute and pumping costs, credit the delivered electrochemical power and the recovered thermal availability, and divide by the useful computational throughput,

$$J_{\text{rel}} = \frac{\displaystyle\int_0^{\tau} \left(P_{\text{comp}} + P_{\text{pump}} - P_{\text{fc}} - \eta_{\text{ex}}\,\dot{X}_{\text{th}}\right) dt}{\displaystyle\int_0^{\tau} \Lambda\, dt}.$$

The claim attached to this objective is comparative, not absolute: under nonzero K1–K3, minimizing J_{rel} over the shared network strictly dominates any policy that optimizes cooling, power delivery, or computation in isolation, because the isolated policies cannot represent the credits and the shared constraint that the couplings introduce. The full decision variables, governing equations, ledger of every term, and constraints are stated in the companion defensive publication; they are reproduced here only far enough to make the argument legible. The interactive companion demonstrates the qualitative consequence: its jointly chosen operating point is an interior optimum — neither the coldest nor the cheapest setting — that a single-current objective does not find.

9. Determinism as a safety floor

An analog fluid computer raises an obvious objection for any system that must be trusted to act in the world: if the answer depends on temperature, wear, and manufacturing spread, it is not reproducible, and a Physical AI whose outputs are not reproducible cannot be verified or certified. The report treats determinism as a hard, non-negotiable constraint rather than a goal

to be traded against efficiency. The mechanism is to commit not to the analog readout but to a quantized word: each node's output is passed through a fixed-point quantizer with a least-significant-bit δ chosen larger than the worst-case analog spread across the entire feasible operating envelope, so that every admissible temperature and flow maps to the identical emitted word. The committed observable is the fixed-point trajectory, and its integrity is checked by hashing:

$$H = \mathrm{SHA\text{-}256}\big(\hat{y}_k(t)\big) \stackrel{!}{=} H^{\mathrm{ref}}.$$

Operating point may change the *speed* and the *cost* of the computation; it may never change the emitted word. This ties the microfluidic substrate to the Institute's broader thesis that Physical AI compute should be joule-priced *and* bit-exact — efficiency measured against a floor, correctness measured against a hash — and it converts the analog nature of the medium from a liability into a bounded, verifiable design parameter. It also closes back onto efficiency: running hotter widens the analog spread, forcing a coarser quantizer, which lowers throughput. Determinism thus prices cooling headroom, giving K2 and K3 a hard edge the optimizer must respect.

10. Maturity: what ships, what is research

Honesty about readiness is the point of this section, because the three currents are separated by many years of development and it would be misleading to present them as equally near. Table 2 gives the assessment.

Table 2. Maturity assessment. The integration is a research programme; two of its three legs stand on mature or demonstrated work, the third does not.

CAPABILITY	STATUS	HONEST CAVEAT
Embedded microfluidic cooling	Engineered; demonstrated at chip scale	Fabrication and reliability of in-silicon channels remain non-trivial
On-chip fluidic power	Lab-demonstrated; combined cooling+power shown	Power density is modest; a supporting role, not primary supply
In-fluid iontronic computation	Early research; single devices and small networks	ms–s timescale; not a processor; durability and scaling unproven
Joint co-optimization (this work)	Position + illustrative model + prior-art disclosure	No integrated device built; the model is first-order, not validated

The single most important caveat is the timescale of the third current. Iontronic devices relax on the order of milliseconds to seconds, which places them three to six orders of magnitude slower than digital logic. Any claim that a fluid could *replace* a processor is therefore wrong, and this report makes no such claim. What the timescale does support is a slow layer — sensor conditioning, homeostatic regulation, reflex-speed pattern detection — that runs where the fluid already is, at a marginal exergy cost near zero because the cooling and power infrastructure would be present regardless. That, and not a faster computer, is the opportunity.

11. Discussion: a research position

The position this report advances is narrow and, we think, defensible. It is not that fluid computers will be fast, nor that they will be general; the physics forbids the first and the maturity forbids the second. It is that an embodied machine must carry a thermal path and a power path, that both are increasingly built as fluids, and that a fluid that is already present can carry a slow computational layer for very little additional cost — provided the three are designed together, priced in exergy, and constrained to a deterministic readout. Each of those three provisos is doing work: designed together, because the couplings make separate optimization wrong; priced in exergy, because energy accounting hides the recovery that makes the scheme attractive; deterministic, because an embodied AI that cannot reproduce its own outputs cannot be trusted with a body.

The work is deliberately released as an open commons. The constituent physics — Poisson–Nernst–Planck transport, Butler–Volmer kinetics, the Soret effect, exergy accounting — is entirely public, and the contribution here is the framing and the joint objective, which we place in the public domain as prior art through the companion defensive publication rather than enclose. The intended readers are the three communities that have been working apart: this report is an argument that their problems are, on an embodied device, one problem.

12. Conclusion

An embodied machine's computer lives inside its cooling and its power supply, and both are increasingly fluids threaded through the active silicon. The same electrolyte can hold and read information at nanoscale nodes, adding a third current to a network that must exist for the first two. Because the three share one flow and one dissipation budget and are coupled by ordinary physics, they should be optimized jointly and priced in exergy, and because the machine must be trustworthy, the readout should be held bit-exact. The near-term prize is not speed but a slow homeostatic layer obtained almost for free in the infrastructure an embodied system already carries. The claim is a research position, grounded in mature cooling, demonstrated fluidic power, and early but real iontronic computation, and it is offered as an open commons for the fields it asks to be read together.

References

1. R. van Erp, R. Soleimanzadeh, L. Nela, G. Kampitsis, E. Matioli. Co-designing electronics with microfluidics for more sustainable cooling. *Nature* 585, 211–216, 2020. doi:10.1038/s41586-020-2666-1.
2. M. Zhang, J. Zhu, J. Li, H. Feng, D. Hu, X. Chen, Q. Li. Chill and charge: a synergistic integration for future compact electronics. *Device* (Cell Press) 2, 100367, 2024. doi:10.1016/j.device.2024.100367.
3. J. W. Lee, M.-A. Goulet, E. Kjeang. Microfluidic redox battery. *Lab on a Chip* 13, 2504–2507, 2013. doi:10.1039/c3lc50499a. (Membraneless co-laminar flow cell.)
4. L. O. Chua. Memristor — the missing circuit element. *IEEE Trans. Circuit Theory* 18(5), 507–519, 1971.
5. T. M. Kamsma, W. Q. Boon, T. ter Rele, C. Spitoni, R. van Roij. Iontronic neuromorphic signaling with conical microfluidic memristors. *Phys. Rev. Lett.* 130, 268401, 2023. arXiv:2301.06158.
6. T. M. Kamsma, J. Kim, K. Kim, W. Q. Boon, C. Spitoni, J. Park, R. van Roij. Brain-inspired computing with fluidic iontronic nanochannels. *PNAS* 121(18), e2320242121, 2024. doi:10.1073/pnas.2320242121.

7. T. M. Kamsma, J. J. Teijema, R. van Roij, C. Spitoni. Echo state and band-pass networks with aqueous memristors: leaky reservoir computing with a leaky substrate. *Chaos* 35, 093133, 2025. arXiv:2505.13451.
8. S. R. de Groot, P. Mazur. *Non-Equilibrium Thermodynamics*. North-Holland, 1962 (Dover reprint 1984). (Coupled transport; the Soret / thermodiffusion effect.)
9. A. Bejan. *Advanced Engineering Thermodynamics*. Wiley, 4th ed., 2016. (Exergy / availability accounting.)
10. Embedded microfluidic cooling with compact double-H-type manifold microchannels for large-area high-power chips. *Int. J. Heat and Mass Transfer* 197, 123340, 2022. doi:10.1016/j.ijheatmasstransfer.2022.123340.

AI-use disclosure. Preparation of this report used a large language model (Claude, Anthropic) for drafting and editing text, organizing the reviewed literature, and preparing the figures and the interactive companion. Cited references were checked to resolve to their sources. The authors reviewed the content and are solely responsible for it. Consistent with ICMJE, COPE, and IEEE guidance, the model is a tool and is not credited as an author.

Institute for Physical AI @ BMI · The Charlot Lab & The Hiner Lab
503 McKeever Rd, Arcola, TX 77583, USA
physicalai-bmi.org · contact@physicalai-bmi.org

Technical Report TR-2026-04 · Preprint v1