I. Benefits of Cooling

In cryogenic systems, the Coefficient of Performance (COP) is given by Q/W, where Q is the heat removed from the system and W is the work done by the refrigerator. The maximum possible COP is determined by the Carnot Efficiency and is given by:

$$COP_{optimal} = \frac{T_L}{T_H - T_L} \tag{1}$$

For a device operating in a cryogenic environment, the total power consumed is the sum of the power used by the device (P_{Device}) and the power consumed in removing the excess heat (P_{Device}/COP). The total power is then:

$$P_{tot} = (1 + \frac{1}{COP})P_{Device} = \frac{T_H}{T_L} P_{Device} \tag{2}$$

It is common for P_{Device} to be a function of temperature. If P_{Device} goes as T^{α} then the ratio of power consumed at T_L and T_H is:

$$\frac{P_L}{P_H} = (\frac{T_L}{T_H})^{\alpha - 1} \tag{3}$$

If a device is to ultimately require less power to operate cold than warm, then P_{Device} must scale at least linearly with temperature. This has important ramifications for the fundamental limits of computation. For instance, the Landauer Limit states that the minimum energy to either generate or erase a bit of information is:

$$E_{min} = kT \ln 2 \tag{4}$$

Sometimes the Landauer Limit is stated to be 2.805 zJ at room-temperature. An even stronger statement would be to say that the Landauer Limit is 2.805 zJ *on earth's surface* (exempting Antarctica). Since the Landauer Limit only scales linearly with temperature, in terms of total power, no refrigeration method can ever succeed in reducing it below the ambient value.

$kT \ln 2$ also happens to be the minimum possible energy to transmit one bit of information in a noisy channel according to Shannon's Thereom for channel capacity. Once again, refrigeration will be of no use in circumventing this fundamental limit.

For these fundamental limits with linear temperature scaling, the ambient temperature is the ultimate limiting factor. The coldest possible ambient temperature is the Cosmic Microwave Background radiation at around 3K. Thus there are 2 orders of magnitude to be gained in loosening the Landauer and Shannon Limits, but only in some future space-based computing system.

Refrigeration does, however, have the potential to lower the fundamental limit for communicating a bit via charging a physical wire (the Shannon Limit discussed earlier doesn't assume any specific physical mechanism). The energy to charge a wire is CV^2 and the minimum detectable voltage to communicate a bit will be on the order of the thermal voltage. Therefore, the minimum energy to communicate a bit along a charged wire is:

$$E_{wire} > C(\frac{kT}{e})^2 \tag{5}$$

Here the energy scales as T^2 so refrigeration does at least have the theoretical possibility of reducing the total energy to charge the wire.

Lastly, for a more immediately practical example, reverse bias current through a pn junction is an exponential function of temperature: $I_{Leak} = I_s(e^{\frac{-qV}{kT}} - 1)$. Clearly, with such a strong temperature dependence, refrigeration has the potential to reduce the static power consumed by reverse-biased diodes.

[1]https://www.intechopen.com/books/ict-energy-concepts-towards-zero-power-information-and-communication-technology/minimum-energy-of-computing-fundamental-considerations

II. Zero-Amplifier Receiver

As discussed in [Miller], receivers have been proposed without amplification stages in order to reduce power for intra-chip optical interconnects. In this simple configuration, a photodiode transforms photons into electrons that are then stored on a capacitor. This capacitance is the sum of the photodiode capacitance and the input capacitance to a MOSFET. Initially, we will assume no leakage off of this capacitance.

In order to reach the MOSFET's threshold voltage V, we need N_e electrons:

$$N_e = \frac{CV}{q} \tag{1}$$

These N_e electrons are generated from the photodiode absorbing N_p photons. For a detector with quantum efficiency η,

$$N_p = \frac{N_e}{\eta} = \frac{CV}{q\eta} \tag{2}$$

This corresponds to an energy E_{opt} needed to switch the transistor:

$$E_{opt} = h\nu N_p = \frac{h\nu CV}{q\eta} = \frac{CV}{\mathcal{R}} \tag{3}$$

Where $\mathcal{R} = \frac{q\eta}{h\nu}$ is responsitivity of the photodiode.

Aggressive, but plausible numbers may be $C = 1\text{fF}$, $V = 1\text{V}$, and $\mathcal{R} = 1\text{A/W}$. In this case, $E_{opt} = 1\text{fJ}$.

While such a small spiking energy is enticing, the previous calculation completely ignores any time dependence. The charge on the capacitance will invariably be leaking off through some shunt resistance, demanding that the necessary N_p photons must reach the receiver within some fixed period of time. Assuming a shunt resistance R, a square pulse of optical power with magnitude P_{opt} will charge the voltage on the capacitance according to:

$$V(t) = \mathcal{R}P_{opt}R(1 - e^{-t/\tau}) \tag{4}$$

Where $\tau = RC$. Solving for t, we arrive at the neccessary spike duration for a given output voltage and the total energy per spike, E_{opt}:

$$t = -\tau \ln(1 - \frac{V}{\mathcal{R}P_{opt}R}) \tag{5}$$

$$E_{opt} = tP_{opt} = -P_{opt}\tau \ln(1 - \frac{V}{\mathcal{R}P_{opt}R}) \tag{6}$$

Immediately, we see that P_{opt} must be greater than $\frac{V}{R\mathcal{R}}$ for the receiver to ever produce a voltage of V. In the limit $P_{opt} >> \frac{V}{R\mathcal{R}}$, E_{opt} goes to the constant value of $\frac{CV}{\mathcal{R}}$ as we calculated earlier. A plot is shown below with $C = 1\text{fF}$, $V = 1\text{V}$, and $R = 1\text{G}\Omega$.

[[Pasted image 20201020211437.png]]

With these conditions, E_{opt} quickly becomes constant once P_{opt} becomes greater than 1 nW - the minimum possible optical power to reach 1 V. The necessary duration of pulse, however, remains inversely proportional to P_{opt} across many orders of magnitude. In order to maintain 10 MHz frequency of operation, t must be less than 100 ns. This requires P_{opt} to be at least 10 nW. At 1GHz, we'd need 1uW.

Thus far, we have only considered the charging of the capacitance. Any receiver of this type must also have a way of resetting the voltage. In principle this could be achieved by waiting for charge to bleed off through the shunt resistance. However, this introduces a bandwidth/power trade-off. The circuit described above has an f_{3db} of 1 MHz. This problem can potentially be solved with an optical reset of the type desribed in [Miller 2], at the cost of an extra photodiode per synapse.

The question now becomes if a single light source can provide enough power for hundreds or thousands of synapses. At 10 nW per synapse, we would like a light source capable of producing optical powers in the 10-100 μW range. [Romeira] claims that 20 nW is state of the art for on-chip sources right now. If more powerful light sources cannot be developed, neuromorphic networks utilizing PIN detectors will have to investigate other options:

- Returning to a TIA approach in order to detect lower light levels. This is likely tied to an energy budget of 100s of fF/bit. SNSPD detectors would then provide 2 orders of magnitude lower energy per bit - even after acounting for cooling overhead.
- A less elegant design where every synaptic connection is associated with a unique light source. This would at a minimum cost area and complicate routing. It would also increase capacitance on the transmitter side, likely causing delay.

[Miller 1] https://ee.stanford.edu/~dabm/448.pdf

[Miller 2] https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5359/0000/Receiverless-detection-schemes-for-optical-clock-distribution/10.1117/12.518316.short?SSO=1

[Romeira]https://ieeexplore.ieee.org/document/8713372