Advait Comments:

* How to understand this information, such as the key tradeoffs involved and simple first order estimations of the expected performance of such a system would be very beneficial.
  + This was sort of the goal of the paper, although it’s difficult to estimate the expected performance of technologies that are far from mature. We attempt to provide references for the current state of performance of various devices and provide targets for further development.
* how such a system would compare to state-of-the-art neuromorphic platforms/systems (similar scale as this work) such as Loihi, TrueNorth, Spinnaker etc. would in my opinion significantly strengthen the paper. I would suggest adding a table which would report optimistic estimates for semi- and/or super-conducting technologies, with some assumptions for memory and compute circuits. Estimated characteristics such as technology node, core area, learning rules, neuron and synapse density, synapse and neuron operation energies, and estimated speed/time constant would be useful
  + I agree that a thorough comparison of fully-dedicated optoelectronic systems and more popular digital systems is needed. We’ve talked at length about this but think that it is probably best left for an entirely different study. Comparison is not super straightforward, as performance metrics are functions of network topology and size in the digital case. Additionally, many of these systems are highly programmable (their biggest advantage over our proposal), but at the cost of a scale/model complexity tradeoff. No system has yet been built at the scale we are considering (Spinnaker is close at ~1 billion neurons, but only for the simplest neuron model) so we would need to project. Being rigorous about analyzing all these issues feels like a significant undertaking outside the intended scope of this paper.
* The idea here is that if industry is going to want to pursue something of such a scale, 10 to 15 years down the line, the technology has to be competitive with existing systems.
  + Totally agree, but in this work we mostly just wanted to analyze two possible paths toward constructing optoelectronic networks. We refer to our three main constraints as conjectures and hope that future papers will be able to justify them in comparison to current technology.
* Isn’t it true that “Which of the k synapses are going to be connected to each neuron?” is subject to change and plasticity from the learning algorithm. Or do you pre-determine the connectivity and hierarchy?
  + I think this is the primary weakness of our hardware compared to digital systems. The network connectivity and hierarchy must be largely fixed in hardware in this case. *Amend text to make this point clear*
* Building all-to-all connectivity at the 106 neuron scale may be prohibitively expensive and redundant.
  + Yes, almost certainly a non-starter. You need to choose some connection scheme and hope that weights can be learned for that system that solve the problem.
* *Section 2.1.1: Paragraph 3: “We assume that the detection of a single photon will be treated as the registering of a spiking event.”* Is this a reasonable assumption?
  + We think this is reasonable in the superconducting case where SNSPDs have demonstrated extremely low dark count rates (less than 10/sec). I’ve changed the section title from “SOENs Receivers” to “Superconductor Receivers “to make sure readers understand that that statement is limited to the superconducting case.
* How is the memory hierarchy of such a system designed? Is there buffering for local data storage?
  + We restricted the conversation to analog memory integrated right at each synapse, but I do not think we were very clear about this restriction, *so I tried to improve that*. I don’t think that memory hierarchy or buffering are relevant in that case.
* What is the energy cost of an always-on, on-line learning based memory?
  + This is an excellent question, and I believe we referred to this as “peripheral circuitry” throughout the memory section. For the superconducting case, there’s no static power associated with keeping learning mechanisms always online. For the various room-temperature solutions we discuss it would vary, but all of that static power would need to be included when comparing to our benchmark of 3 pJ in table 1.
* In a fully connected approach, with 106 neurons per plane, and 104 planes, how will the all-to-all network be implemented? What is it’s hierarchical structure and what are its performance metrics?
  + *I added a reference to* <https://aip.scitation.org/doi/10.1063/1.5096403> where an all-to-all wiring diagram and a hierarchal network structure is discussed. As I said earlier, all-to-all can’t be implemented at the 10^6 neuron scale, but it might useful locally, and at least provides an example before a specific connection scheme is chosen.
* How much area on each electronic plane is going to be dedicated to the network?
  + Very little, communication is totally passive once optical spikes are produced so there’s no multiplexing/arbitration electronic overhead.

Alex Comments:

* Conjecture 2 – spiking signals – rules out most artificial neural networks used today (except neuromorphic CMOS), which use continuous valued signals. I think it is justified by “Further, performing synaptic weighting in the electronic domain allows for binary optical communication, which minimizes the amount of optical energy per spike and reduces noise incurred by communication.” Isn’t this circular logic, if the second part of the sentence is assuming spikes?
  + Conjecture 2 is not meant to justify spikes; spikes are just assumed to be the goal. Conjecture 2 is meant to say something along the lines of, “If you’re making a fully-dedicated spiking network, it is best to use binary optical pulses to communicate spiking events for noise and energy reasons.” This is in contrast to a platform that might attempt to weight signals optically and attempt to communicate the weighted signals. *I changed the intro to make it clear that we are talking about spiking networks from the beginning and that spiking isn’t something we’re trying to justify.*
* Conjecture 1 – dedicated axons – eliminates all neuromorphic electronics and photonics in use today. It could use a sentence about advantages of “fully dedicated axon approach” after citing Ref. [4]. Reviewers and readers might argue that multiplexing makes sense for the brain-scale goal because it trades of bandwidth (abundant) for interconnection density. Would it be possible to rebut that argument ahead of time?
  + *I added a sentence here to emphasize that multiplexed systems will run into bottlenecks*. We think that a thorough study of when full dedicated systems outperform multiplexed platforms is worth an entire paper on its own.
* It would help to distinguish these between hard to develop and fundamentally impossible, if you can, possibly in the table in Fig 7.
  + I’m not sure anything we discussed is fundamentally impossible.
* What is your opinion of the fruitfulness of this approach? I walk away with an impression that the prospect of brain-scale neuromorphic hardware is pretty bleak, especially if it is guided by conjectures 1-3. It is a lot of very difficult todos. Do you think it is actually possible? If so, then that should be prominently featured in the abstract and conclusion.
  + I don’t think there’s anything that appears outright impossible for either case, but we also enumerated a lot of undeveloped things that must break our way. The chance of all of them doing so is probably not high. I’m not sure that the situation improves much by abandoning conjectures 1-3 – A large, slow system with an electronic digital communication infrastructure and floating gate memories might be possible, but I’m not sure how much speed you can afford to sacrifice before getting beat out by plain old digital neural simulations.
* When referring semiconductor memory technology, I think it would help to clarify that *analog* memory is being stipulated.
  + Yes, the other review had a similar issue. *I’ve changed that section to make it clear that we are talking about local, analog memory.*
* Fig. 6 might be better as an equation. Example of graphical information would be data points on this, such as modern CMOS neuromorphic (1M, 1kHz), modern RT neuromorphic photonic architectures (10, 1GHz), memristors (?, ?), plus future CMOS (?M, 1kHz), future RT photonic (10k, 40GHz), plus some of the typical operating points you expect from the systems analyzed in this paper.
  + *I added the equation in the caption so people can see the relationship between variables.* I think that the figure is still sort of nice for seeing specific cases though. I’m a little hesitant to add data points for other technologies though. Many are at scales that are far smaller than what we’re discussing and I’m not sure about attempting to project without a whole lot of knowledge about those systems.
* Sec 5.5.2. Why choose 10MW? If you can assume that superconductor-semiconductor interface is possible (big “if”), can’t you assume that GW computers can be made?
  + I suppose, although I think it’s actually more encouraging to show that you might be able to make some really high-performance systems even without assuming any increase in available power.
* Multiple places, there is reference to “Table 4.1.4” Do you mean table 1?
  + *Yes, thank you.*
* Circuit diagrams: put dots on top of wire crossings to indicate where there is a junction between wires
  + *Done*.