Rich Comments:

* Pg. 1 “require III-V light-source integration with electronics at an unprecedented scale” – is this unprecedented, what about LUMOS?
  + I think so, LUMOS doesn’t seem as focused on electronics aspects. Does it even specify CMOS integration?
* Pg. 3 “light-sources – a speculative element…” – Avoid commentary
  + Deleted commentary, changed to: “However, the optical power required from light sources is reduced by a factor of 1000…”
* Pg. 3 “Lspd can be as low as 100 nH…” – Is this a realistic value?
  + If we take the value of 160pH/square from MoSi, this allows us to have 625 squares. If the nanowire is 100nm, this allows 62.5um length, which gives 20 dB of attenuation, which allows the 99% detection consistent with detecting 99% of spikes. Actual numbers may vary by about a factor of 2 depending on material, wavelength, waveguide structure, etc, but that’s okay for now.Pg. 3 “Detecting only 99% of spikes may be tolerable” – Is this a useful value?
* Yes, it is. Biological synapses are far less reliable than even 99%. I’ve added the following reference to support this: <https://www.pnas.org/content/91/22/10380.short>

And changed the text to: “Detecting only 99\% of spikes may be tolerable and would still represent a significant improvement over biology, wherein synapse reliability is typically in the range of 5\% - 80\% \cite{allen1994evaluation,li1997}}”

* Pg. 5 In regards to III-V integration, change aspirational to highly challenging
  + Done
* Pg. 5 “Epitaxial growth would be the optimal solution” – Not necessarily optimal for LEDs
  + Changed to: “Epitaxial growth would be an attractive solution due to the high throughput…”
* Pg. 5 “the semiconductor platform would require more expensive and less scalable bonding technologies.” – Why is this less scalable?
  + Changed this to: “the semiconductor platform would be less scalable due to the limited size of III-V wafers and the expense of performing wafer bonding.”
* Pg. 6 Change “fj to pj” to “femtojoule to picojoule”
  + Done

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. Do you have recommendations for specific performance specifications you would like to see?
* 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 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. Actually we would go a step further and say being competitive with existing systems is insufficient. This hardware must outperform existing systems by at least an order of magnitude in key metrics if it is going to be adopted. 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 our previous studies as well as future papers justify these conjectures 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. As we envision learning, plasticity modifies synaptic weights and neuronal thresholds, but does not result in adding or eliminating axons. While highly reconfigurable networks are useful for research purposes, such reconfigurability inevitably brings hardware overhead. Tailoring the network to a specific application brings performance benefits. I added a line in the introduction to try and make this clear from the start:
    - “While this strategy requires largely fixing network topology in hardware --- a chief disadvantage when compared with highly reconfigurable digital systems --- the reduced overhead and elimination of communication bottlenecks will greatly benefit performance.”
* 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. We are not aware of any networks in the brain that utilize all-to-all connectivity. Sparse connectivity is observed in the hippocampus and neocortex. There’s quite a bit of theoretical work around the idea that short network path length is necessary for cognition, which is why we based all of the scaling analysis in the paper around random graphs with connectivity fraction set by the average path length.
* *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. I don’t think that memory hierarchy or buffering are relevant in that case.
    - I modified the intro paragraph to this section to include the following line: “A local, analog memory element unique to every synapse will provide the most efficient performance by eliminating memory retrieval and digital conversion.”
    - Also, in response to this and another comment, I’ve changed the title of the subsection “Room-temperature Technologies” to “Room Temperature Analog Memories”
* 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 that 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.
    - I added the following sentence to make sure this was clear: This value includes any energy consumption of peripheral circuitry, both static and that associated with programming.
* 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 referenceto<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, and even if it could, it would not be advantageous as neurons would receive too much stimulus and would fail to specialize, but it might be useful as a case study, and at least provides an example before a specific connection scheme is chosen. In our previous work we have focused on networks that either have random connections (like the hippocampus, which serves as a reservoir) or power-law connections obeying Rent’s rule (which serve to minimize network path length while keeping wiring costs low).
* 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. I hope that the following sentence will provide clarity.
    - “Photonic planes will implement the passive optical interconnects and electronic planes will accommodate all active electronics for neuronal function.”

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 weight signals optically and attempt to communicate the weighted signals.
  + I’ve changed the intro to state that the goal is “a hardware capable of simulating spiking neural networks with scale and complexity of the brain…” and stated that we’re talking about spiking networks in the abstract in order to make it clear we are assuming spiking from the very beginning.
  + Also, a slight change the statement in question: “Further, performing synaptic weighting and temporal dynamics in the electronic domain allows for binary optical communication, which minimizes the amount of optical energy per spike and reduces noise incurred by communication.”
* 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’ve added the following to try and emphasize the issues with multiplexed communication:
    - “While digital systems partially circumvent this issue by leveraging time-multiplexing to artificially increase fan-out \cite{young2019review}, multiplexing introduces latency that scales exponentially above a certain data load \cite{hennessy2011computer}. Optical interconnects may enable direct connections between neurons which would eliminate all traffic-induced delays and support larger, faster, and more interconnected networks.”
  + We think that a thorough study of when fully 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. If Shainline had to guess, he’d say low-capacitance photodiodes are not impossible (a position strengthened by the fact that they’ve been abundantly demonstrated), while CMOS-integrated III-V light sources may prove prohibitively difficult (a position based on the last thirty years of hearty funding and effort in this area without a single successful demonstration). But these positions are exactly the opposite of the reviewer, which goes to show that it is perhaps unwise to stake firm claims regarding possible and impossible technologies.
* 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. We don’t know if it is possible, and we don’t wish to over-hype or prematurely condemn the concept. We think the demonstrations collected in Fig. 7 elucidate the necessary hardware demonstrations required if brain-scale artificial cognition is to be achieved.
* When referring semiconductor memory technology, I think it would help to clarify that *analog* memory is being stipulated.
  + Yes, the other reviewer had a similar issue.
  + I modified the intro paragraph to this section to include the following line: “A local, analog memory element unique to every synapse will provide the most efficient performance by eliminating memory retrieval and digital conversion.”
  + I’ve changed the title of the subsection “Room-temperature Technologies” to “Room Temperature Analog Memories”
* 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 helpful for seeing specific cases. I’m 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. We have added all suggested data points to the plot. They don’t show up because they’re very far to the left or below the plotted region of interest.
* 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. We have chosen 10MW because it is the power consumed by a typical supercomputer. We are assuming a superconductor-semiconductor interface is possible because we have demonstrated it and published the result in Nature Electronics. Further work needs to be done to ensure the devices reset after pulsing. We do not think it is a comparable assumption to speculate that in the future power consumption will be irrelevant.
* 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
  + While that probably would improve clarity, none of Jeff’s previous papers have used dots to signify junctions, so we’re just being consistent at this point.