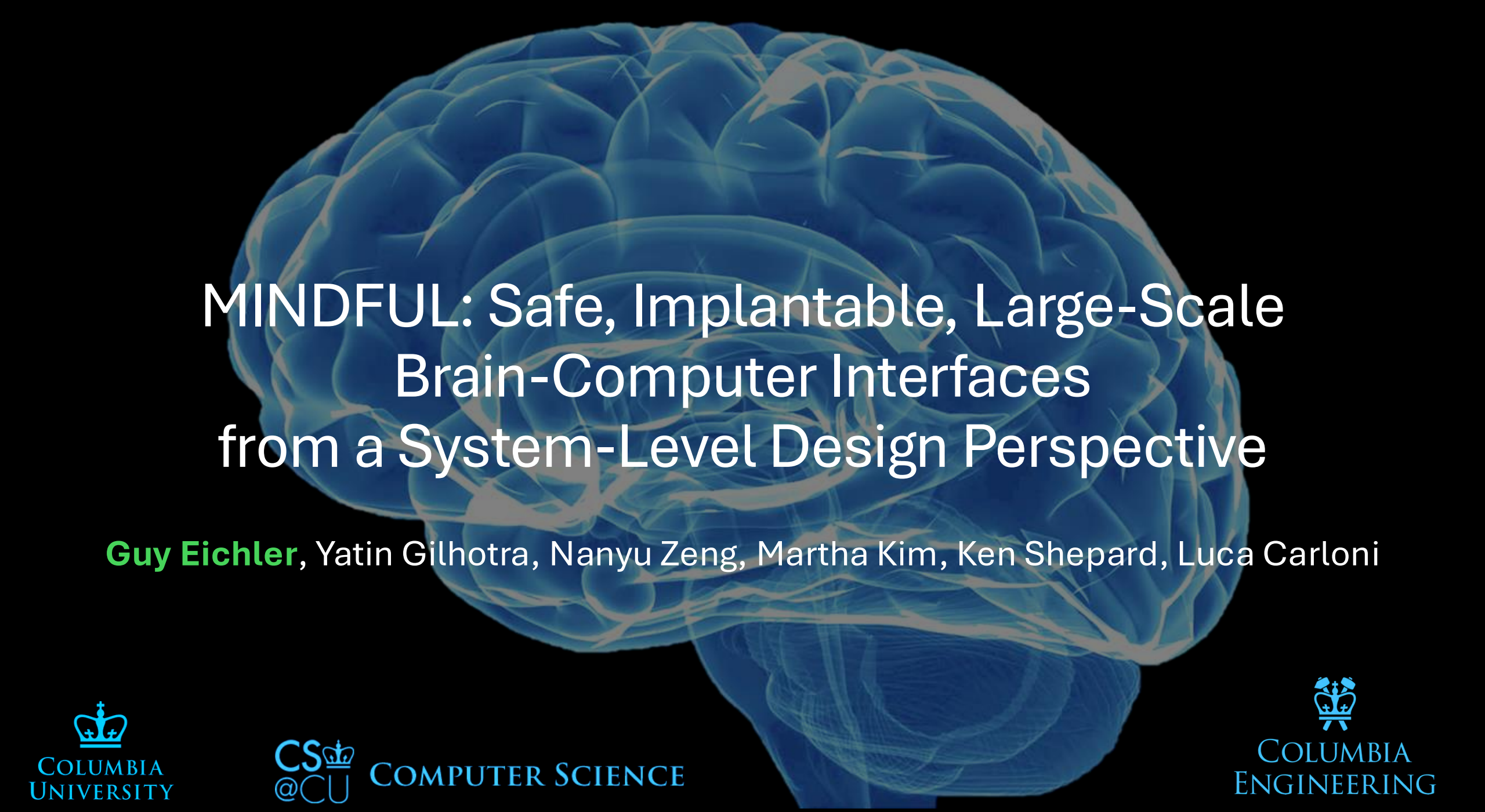


MINDFUL: Safe, Implantable, Large-Scale Brain-Computer Interfaces from a System-Level Design Perspective

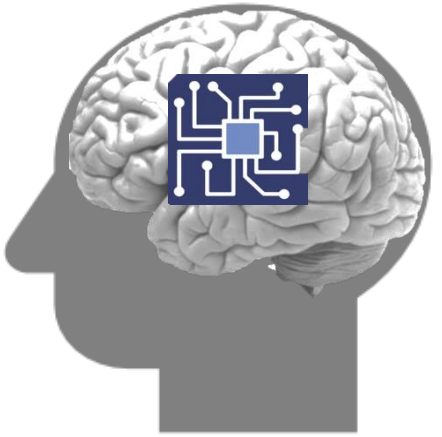
Guy Eichler, Yatin Gilhotra, Nanyu Zeng, Martha Kim, Ken Shepard, Luca Carloni

Columbia University

CS@CU Computer Science

Columbia Engineering

Brain-Computer Interfaces (BCIs)

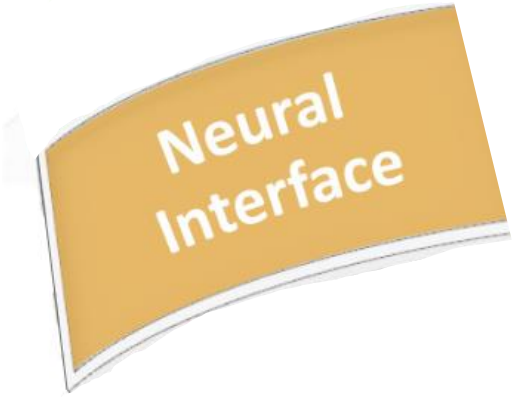


Connecting the brain with the digital world for –

- **Healthcare** - improve quality of life (***online***)
- **Brain research** – understanding the brain (***offline***)

Neural Interface (NI) – a set of ***channels***

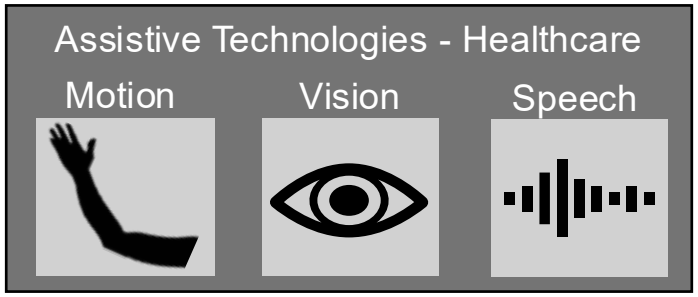
- Non-invasive – EEG, EMG, ultrasound
- **Invasive** – ECoG, intracortical, optical imagers
- Sampling frequency - $f_{sampling}$

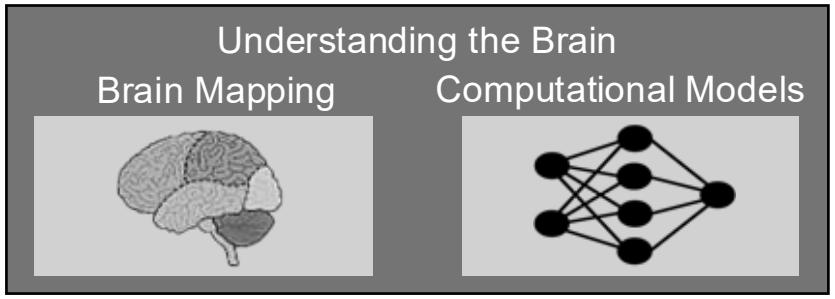


Communication – **wireless** vs. wired

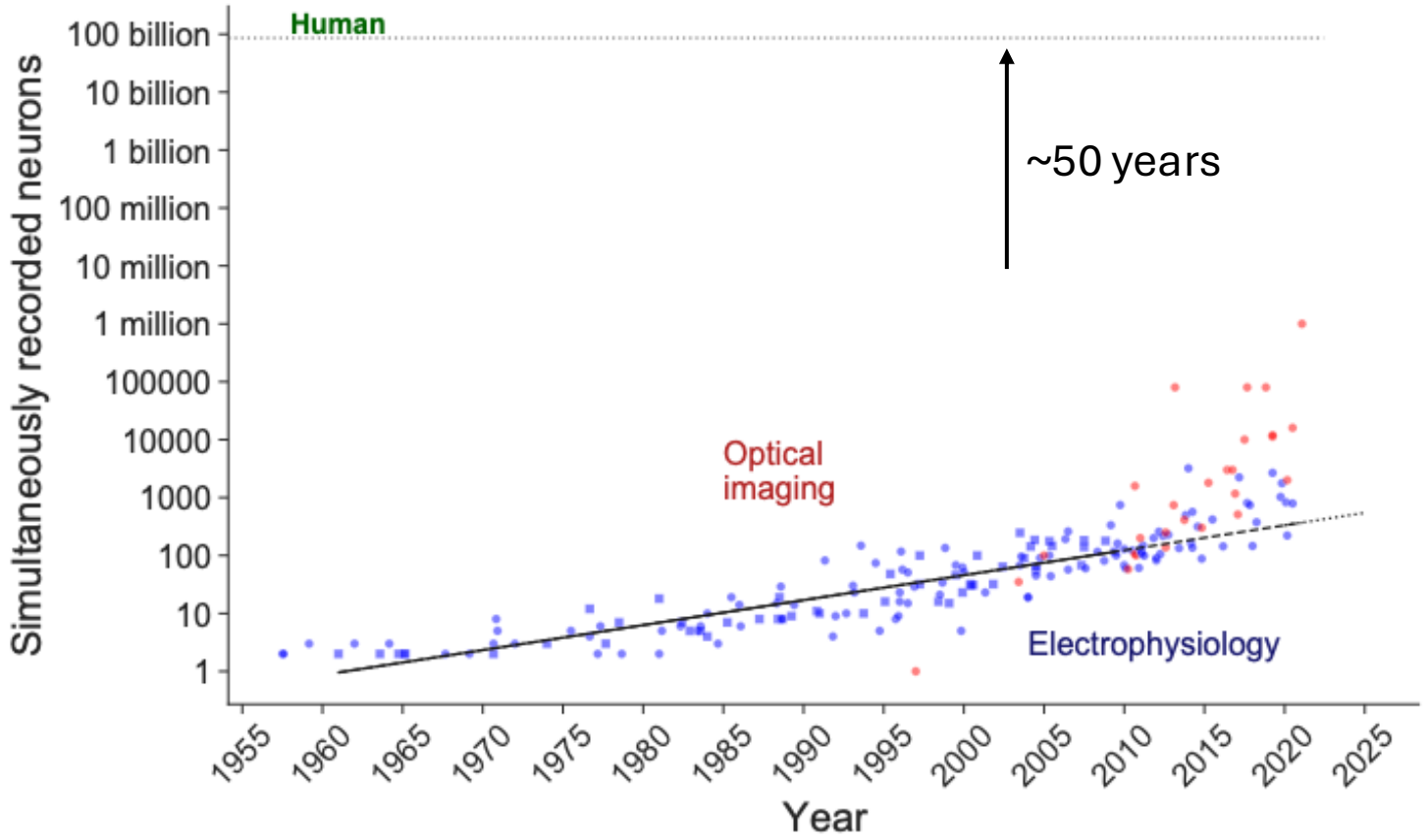
Computation – packetizing vs. application-level

External processing – less constraints





“Moore’s Law” for BCI Design – Neural Interface



How does it affect the BCI system?

Real-time?
Low power?
Computation?
Communication?

We need an analytical framework!

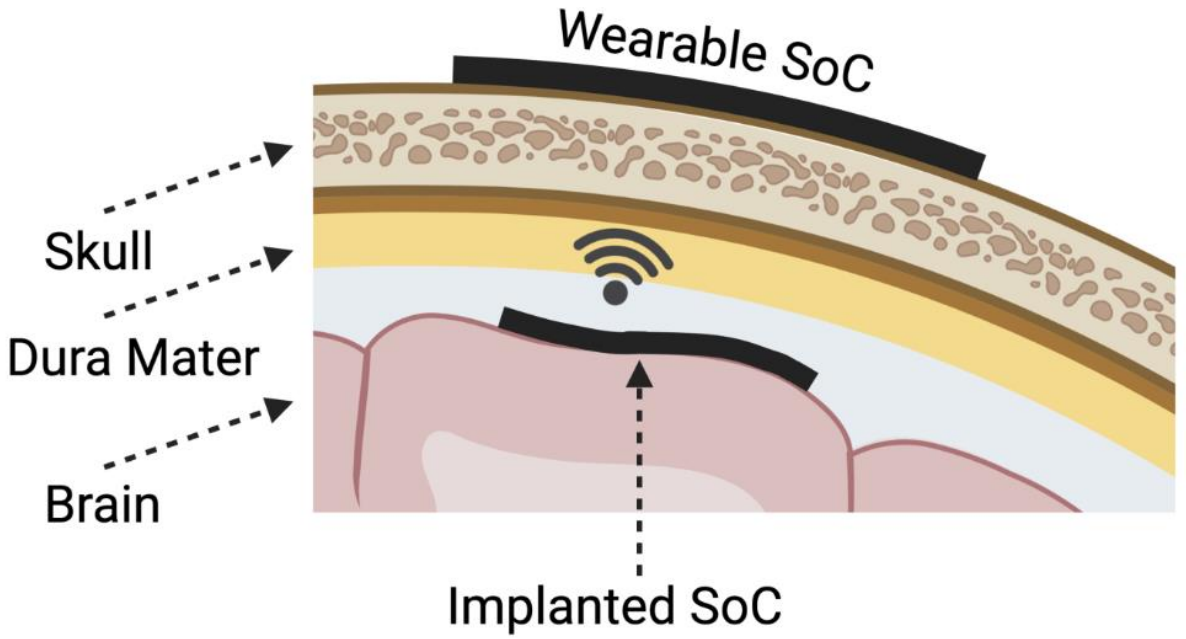
Number of recorded neurons – neural interface size – doubles every 7 years!

[1] “Large-Scale Neural Recordings Call for New Insights to Link Brain and Behavior”, Urai et al., Nature Neuroscience, 2022
[2] “How Advances in Neural Recording Affect Data Analysis”, Stevenson and Kording, Nature Neuroscience, 2011

1. How to design a BCI system?

Target Form Factor and Requirements

- Two-component approach –
  - **Implanted SoC** and **wearable SoC**
  - **Implant** – everything that interfaces with brain tissue
  - **Wearable** – everything that does not
- Uniform heat dissipation: $1 - 2\ ^{\circ}C$
- ***Power density***: $\mathrm{P_d} \approx 40\ \frac{mW}{cm^2}$
- Maximize ***volumetric efficiency***
  - More sensing area vs. non-sensing area
  - Less spacing between channels ($\sim 20 \mu m$)



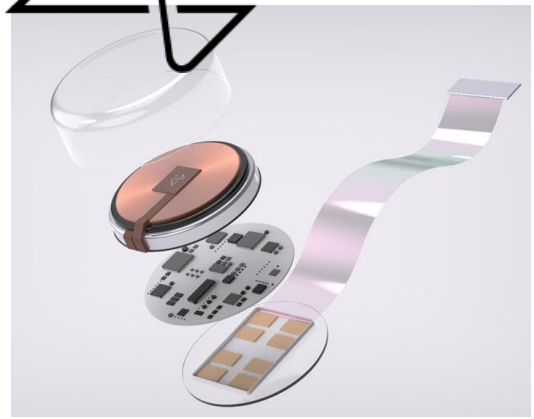
The State of the Art

- Form factor: Implantable and Wireless
- Channels: Electrodes and SPADs (optical imagers)
- Current standard channel count – **1024**
- Most published devices report **low channel counts**
- What should we do? **How do we scale?**

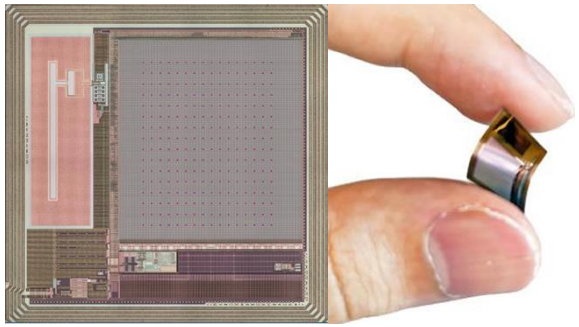
Table 1: Summary of implanted SoC Designs

#	SoC	NI Type	#Channels	Area (mm^2)	P_d ($\frac{mW}{cm^2}$)	f (kHz)	Wireless	In-Vivo/ Ex-Vivo
1	BISC [59, 139]	Electrodes	1024	144	27	8	Yes	Yes
2	Gilhotra *et al.* [42]	SPAD	49152	144	33	8	Yes	Yes
3	Neuralink [85, 136]	Electrodes	1024	20	39	10	Yes	Yes
4	Shen *et al.* [104]	Electrodes	16	1.34	2.2	10	Yes	Yes
5	Muller *et al.* [84]	Electrodes	64	5.76	2.5	1	Yes	Yes
6	Yang *et al.* [132]	Electrodes	4	4	1.3	20	Yes	Yes
7	WIMAGINE [80]	Electrodes	64	1960	3.8	30	Yes	Yes
8	HALO [62, 110]	Electrodes	96	1	1500	30	Yes	No
9	Neuropixels [125]	Electrodes	384	22	21	30	No	Yes
10	Jang *et al.* [56]	Electrodes	1024	3	17	20	No	Yes
11	Pollman *et al.* [89]	SPAD	49152	50	36	8	No	Yes

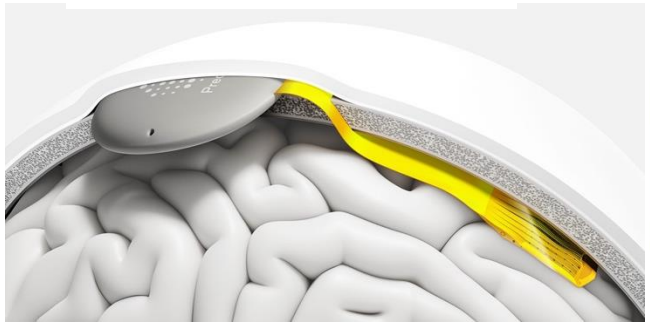
NEURALINK

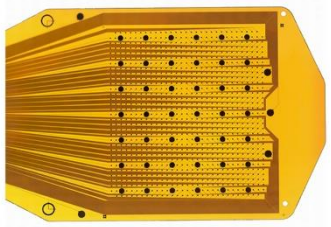


Kampto Neurotech



Precision





2. How to scale a BCI system?

Scaling to the Current Standard (1024 Channels)

Method: Scaling area and power with channel count – **preserve SNR and improve volumetric efficiency**

1. $A_{SoC}(n) \approx \sqrt{n} \cdot A_{channel}$

   $P_{SoC}(n) \approx n \cdot P_{channel}$
2. Scale up **power density** – reduction in area (2x)
3. Reduce area and power – reach reasonable channel spacing for **volumetric efficiency** (~200um)
4. Scale to the closest available **power budget**

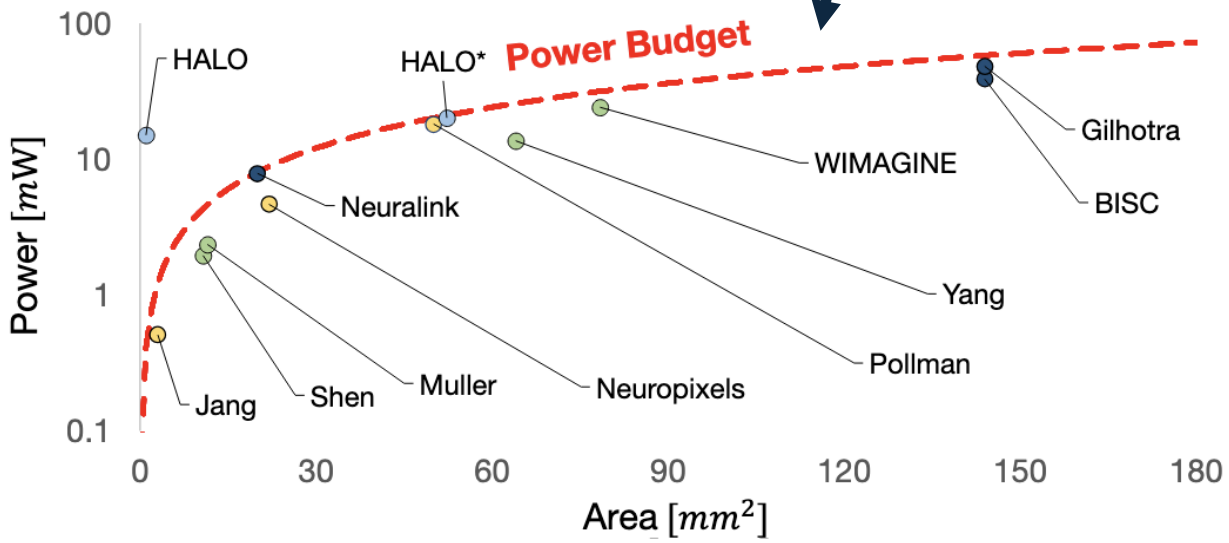
Result: **A set of plausible design points**

(Using the **8 wireless SoC designs** from the literature as our baselines)

***n – number of channels**

Table 1: Summary of implanted SoC Designs

#	SoC	NI Type	#Channels	Area (mm^2)	$P_d(\frac{mW}{cm^2})$	f (kHz)	Wireless	In-Vivo/ Ex-Vivo
1	BISC [59, 139]	Electrodes	1024	144	27	8	Yes	Yes
2	Gilhotra *et al.* [42]	SPAD	49152	144	33	8	Yes	Yes
3	Neuralink [85, 136]	Electrodes	1024	20	39	10	Yes	Yes
4	Shen *et al.* [104]	Electrodes	16	1.34	2.2	10	Yes	Yes
5	Muller *et al.* [84]	Electrodes	64	5.76	2.5	1	Yes	Yes
6	Yang *et al.* [132]	Electrodes	4	4	1.3	20	Yes	Yes
7	WIMAGINE [80]	Electrodes	64	1960	3.8	30	Yes	Yes
8	HALO [62, 110]	Electrodes	96	1	1500	30	Yes	No
9	Neuropixels [125]	Electrodes	384	22	21	30	No	Yes
10	Jang *et al.* [56]	Electrodes	1024	3	17	20	No	Yes
11	Pollman *et al.* [89]	SPAD	49152	50	36	8	No	Yes



[3] "Noise Efficient Integrated Amplifier Designs for Biomedical Applications", Simmich et al., Electronics, 2021.

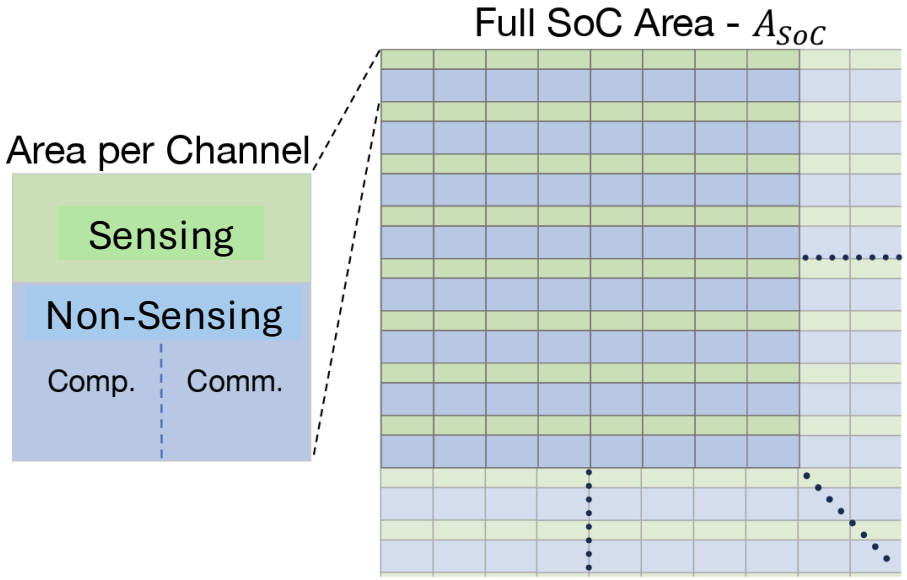
Scaling Beyond 1024 Channels – Next-Gen

Split each SoC into **sensing and non-sensing**:

- $A_{SoC}(n) = A_{sensing}(n) + A_{non-sensing}(n)$
- $P_{SoC}(n) = P_{sensing}(n) + P_{non-sensing}(n)$

Linear scaling for sensing area and power:

- $A_{sensing}(n) = n \cdot \frac{A_{sensing}(1024)}{1024}$
- $P_{sensing}(n) = n \cdot \frac{P_{sensing}(1024)}{1024}$



Non-sensing area and power involves computation and communication – **not necessarily linear!**

Power budget:

- $\frac{P_{SoC}(n)}{A_{SoC}(n)} \leq 40 \frac{mW}{cm^2}$
- $P_{budget} = A_{SoC}(n) \cdot 40 \frac{mW}{cm^2}$

Volumetric efficiency:

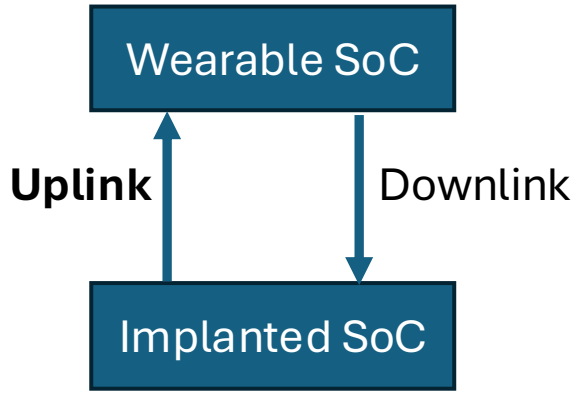
- $\lim_{n \to \infty} \frac{A_{sensing}(n)}{A_{SoC}(n)} = 1$

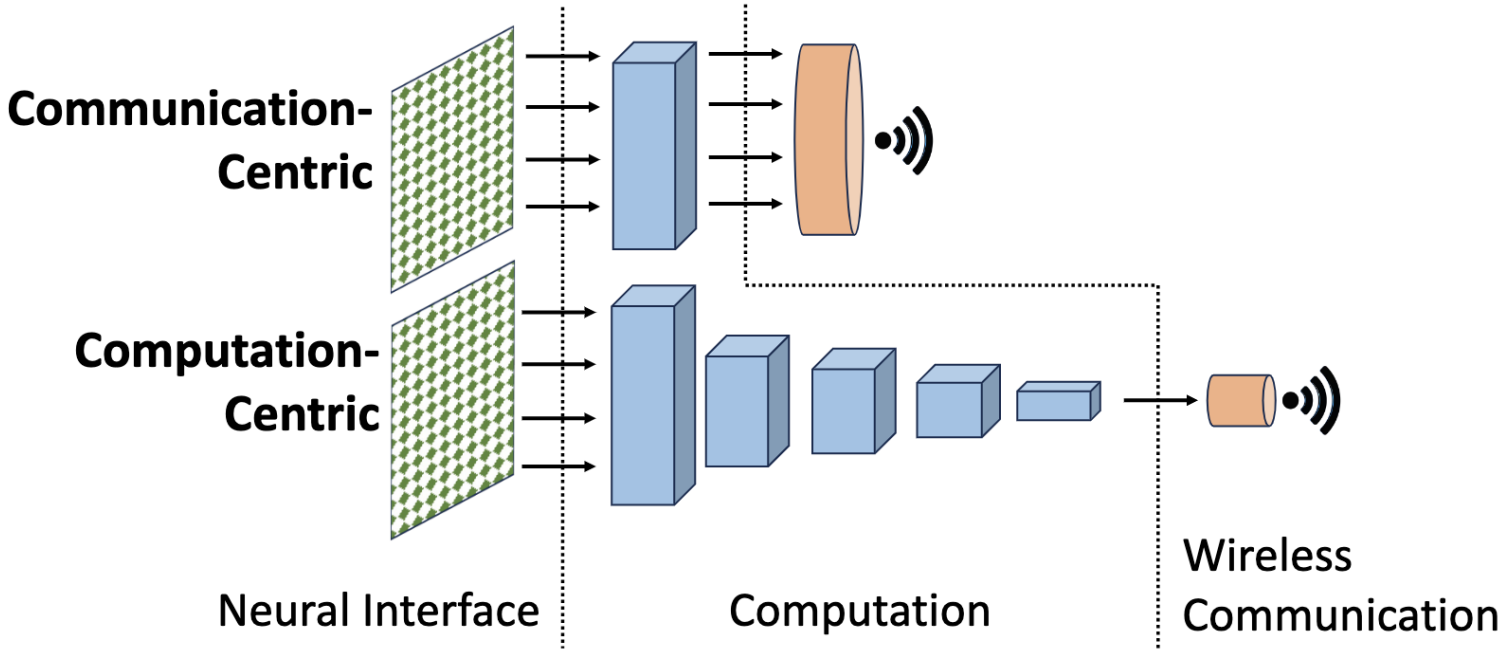
Real-time throughput:

- $T_{sensing}(n) = \frac{d \cdot n}{t}$. $f_{sampling} = \frac{1}{t}$ - NI sampling rate

3. Non-Sensing Architectures

Data Flow





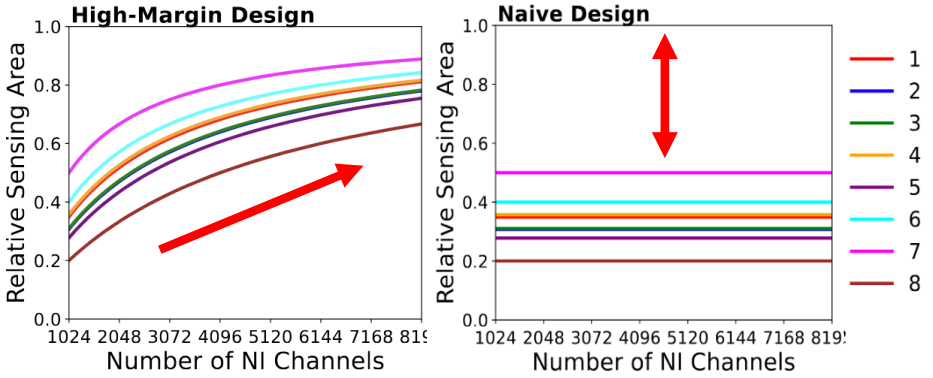
- Focusing on the **uplink** direction
  - Neural data transmission is the dominant operation affecting the bandwidth of the BCI system
  - Controlled by the sampling rate/frequency of the neural interface (NI)
- **Communication-centric** – maximizing data transmission – transmit all digitized neural data recorded on the device
- **Computation-centric** – integrating on-chip application-level computation – send only the "compressed" output

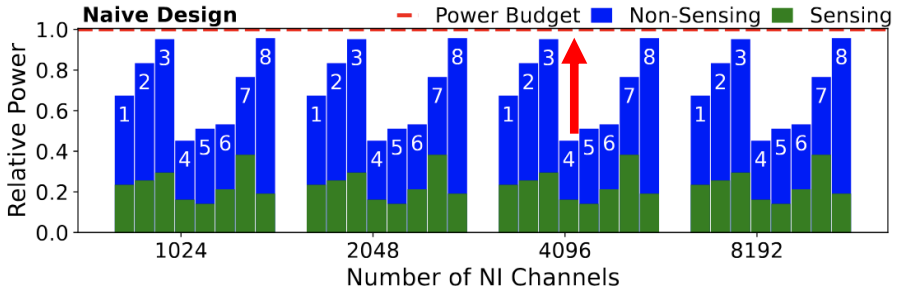
Beyond 1024 – Communication Centric – OOK

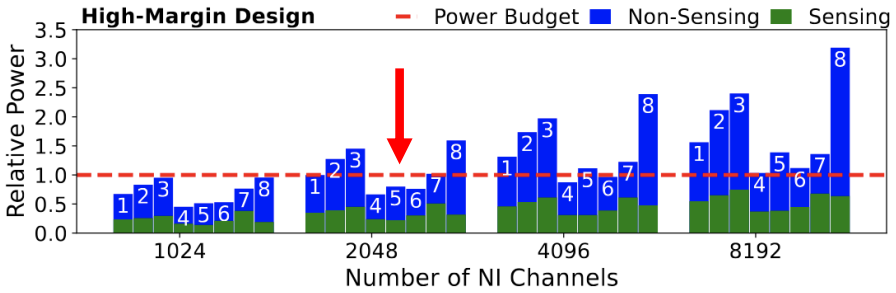
- **Energy-efficient modulation** – On-Off Keying (OOK)
  - Antenna has a maximum BW
  - Transceiver is designed for a maximum data rate
  - Constant energy per bit E_b - 1 bit per cycle

$$P_{comm} = \frac{d \cdot n}{t} \cdot E_b$$

- Experiment: two design strategies –
  - **Naive design** – support higher BW with larger transceiver and antenna (linear increases in area and power)
  - **High-margin design** – original antenna and transceiver can support a higher BW (linear increase in power)
- Two opposing strategies –
  - **Naive** provides optimal power, but area is infeasible
  - **High-margin** optimizes volumetric efficiency, but power is infeasible





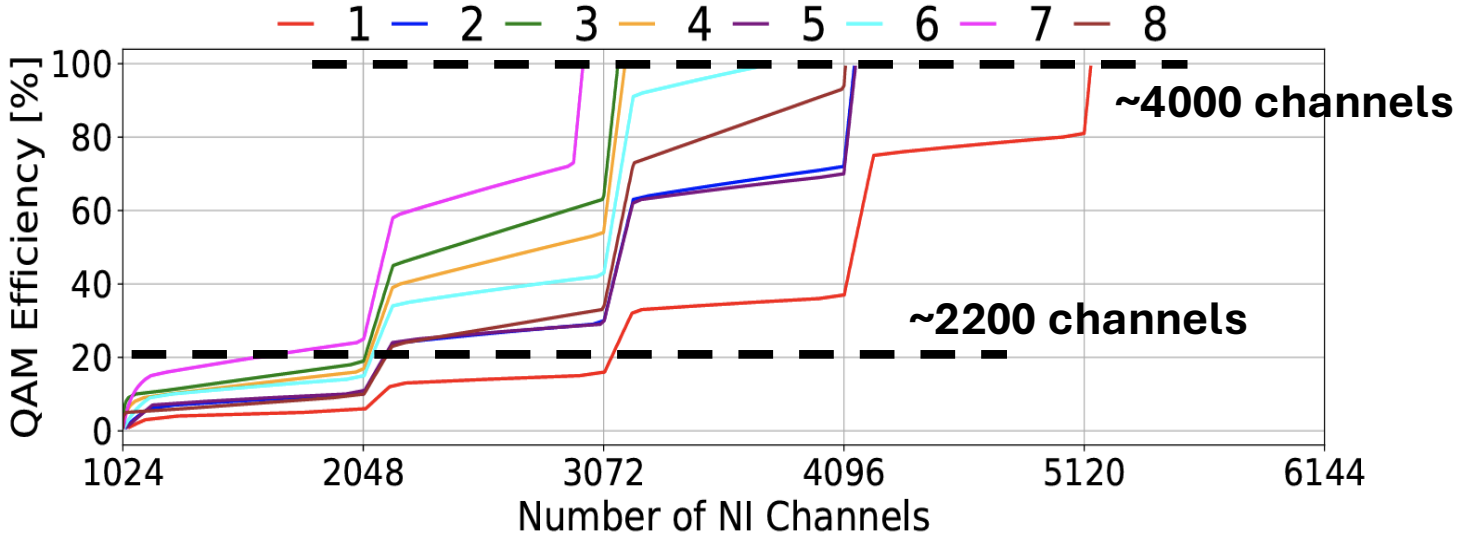


Neither strategy is viable and a hybrid approach is needed to strike a more balanced trade-off between area and power

Beyond 1024 – Communication Centric – QAM

- **Advanced modulation techniques** – Quadratic Amplitude Modulation (QAM)
  - **Advantage:** sending more than 1 bit per cycle/symbol – **higher data rate same antenna**
  - **Disadvantage**: **hard to implement**
  - E_b is calculated for each number of bits transmitted per cycle
- Experiment: measuring **QAM efficiency** – a measure of design quality/effort

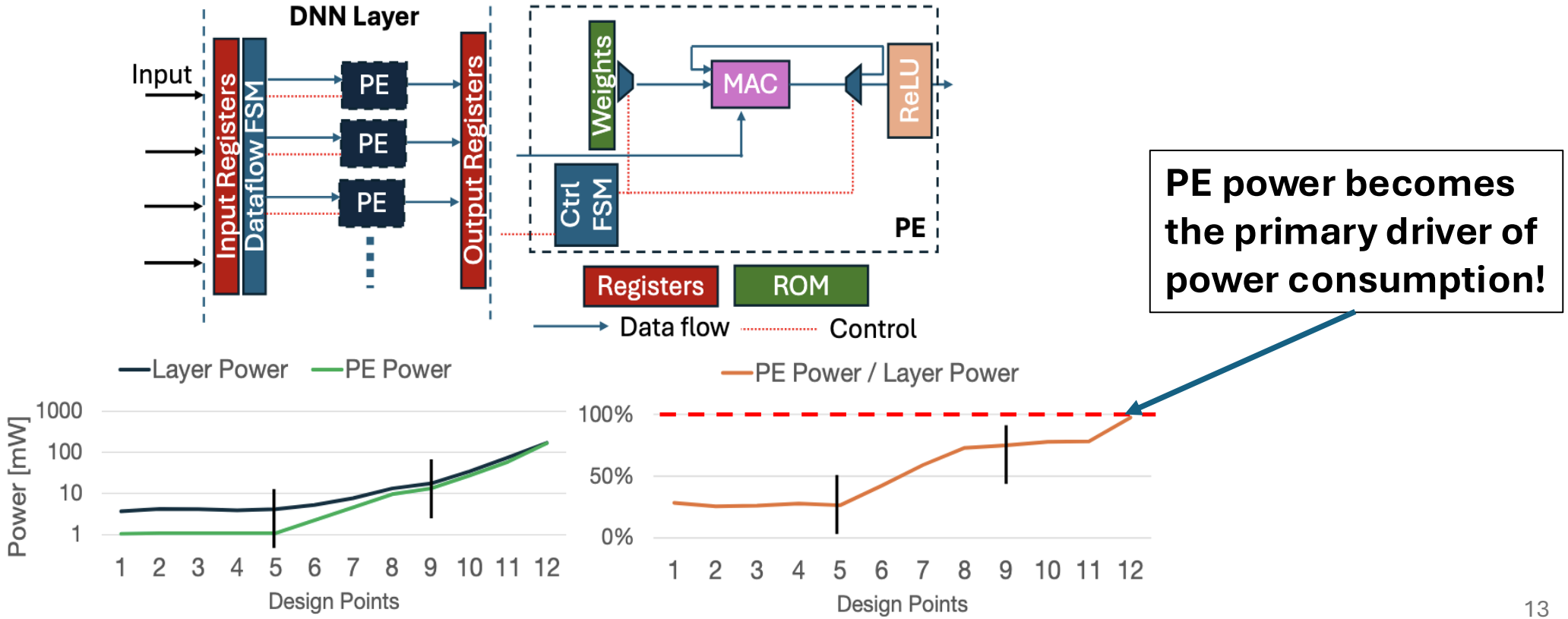
Practical limit of QAM efficiency is usually 15%!



Advanced modulation may meet the higher neural data rates, but it requires overcoming significant design challenges. Eventually, even an ideal yet impractical QAM implementation would not support full neural data transmission, and realistic implementations are likely far more limited

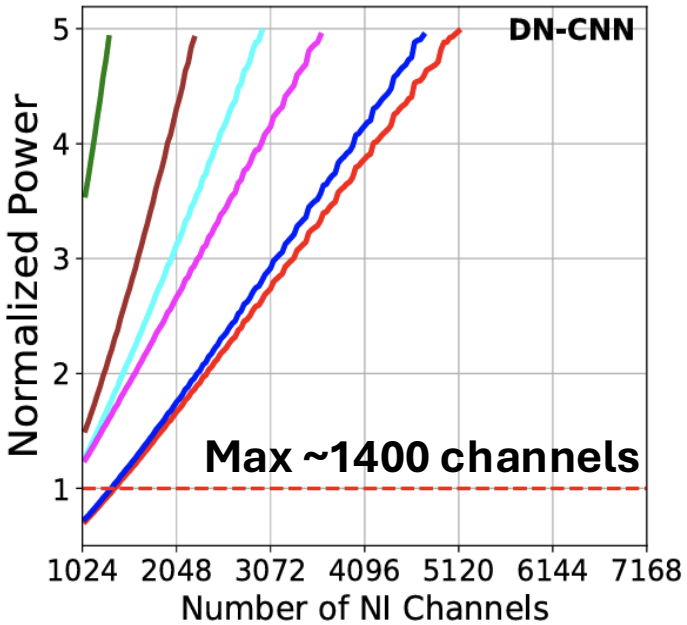
Beyond 1024 – Computation Centric – DNN

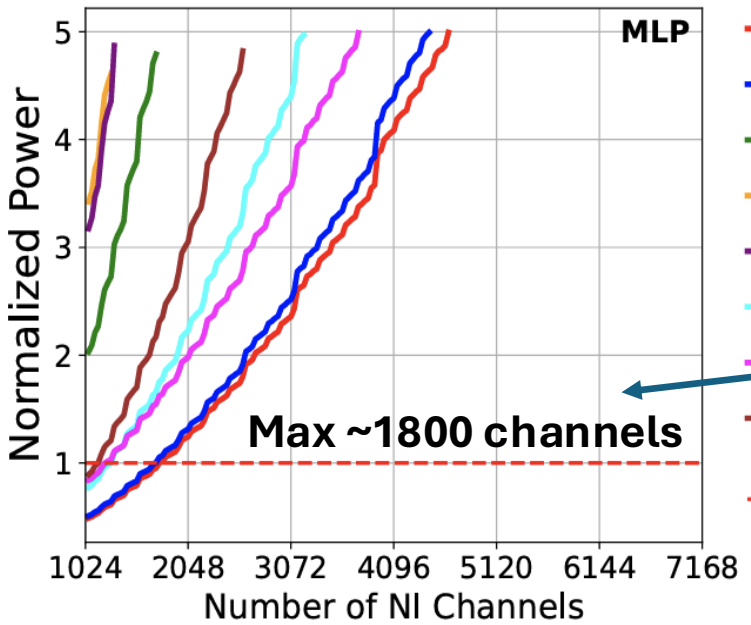
- **Deep neural networks (DNNs)** are currently the popular method in BCI – best accuracy
  - **Multiply-and-accumulate (MAC) units** are the main operation in the DNN architecture
- BCIs implement **non-Von Neumann architectures** – no CPU and main memory – **processing elements (PEs)**
- Experiment: **Scaling** a DNN hardware accelerator

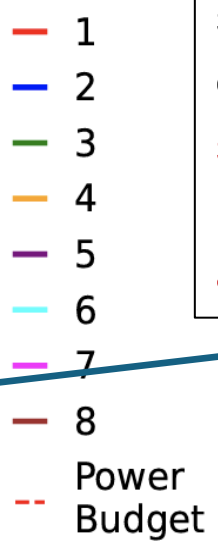


Beyond 1024 – Computation Centric – Scaling

- Experiment: 8 wireless SoCs from Table 1 + 2 DNNs from BCI literature (MLP and CNN)
- **Scale the DNNs with input size (n)** – width and depth
- **Calculate the minimum number of MAC units** to execute the DNN within timing constraints – **solve an optimization problem**
- Get MAC power and MAC time from synthesis results –
  - First-order ***lower bound*** power consumption estimation



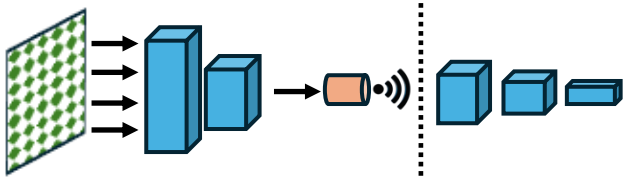




BCI DNNs cannot even scale to twice the current channel count! – raises serious concerns about integrating full BCI applications into BCI SoCs

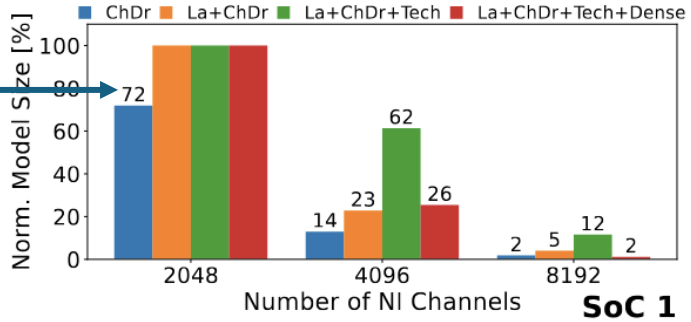
4. Optimizations

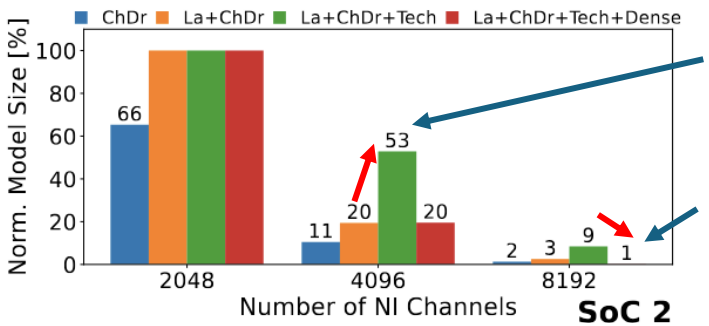
Additional Optimizations - MLP



1. Neural data reduction (**ChDr**) – drop channels to reduce the expected input size and DNN size
2. Layer reduction (**La**) – partition the DNN between the the implant and the wearable
3. Technology node (**Tech**) – upgrade to a newer technology node (from 45nm to 12nm)
4. Channel density (**Dense**) – improve volumetric efficiency by increasing sensing density (2x)

ChDr reduces the DNN model size to fit in the BCI system

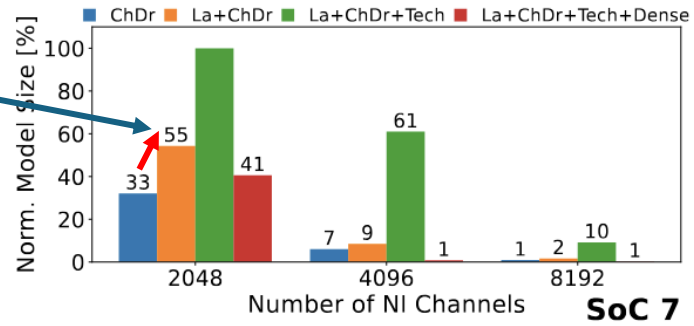


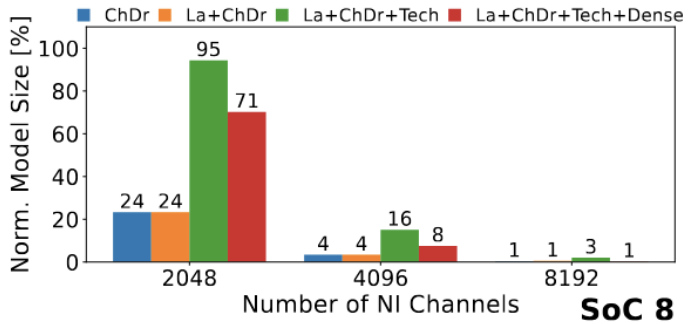


Tech reduces P_{SoC} - a significant increase in the DNN model size

Dense reduces A_{SoC} - lowers the overall P_{budget}

La enables the use of larger DNNs in the BCI system, increasing model size





Reduction in workload size is almost always necessary!

Modern BCI computation should be reevaluated to meet system constraints, while system-level optimizations must adapt to specific applications. Aligning application demands with system limitations is essential for developing safe, implantable, and scalable BCI systems

5. Conclusion and Future Work

MINDFUL is the first analytical framework to support research and development in computer architectures for implantable BCI systems – **open source!**

By establishing a **clearer understanding of the constraints and trade-offs** in BCI system design, we hope to help shape future research directions and foster the development of **safe, scalable, and application-aware BCI architectures**.

List of future work topics includes but not limited to:

NI Scaling Acceleration expected to grow more than double every 7 years!

Recalibration of embedded computation computation must evolve/adapt to changes in brain tissue

Limitations of scaling analog parts analog components do not benefit much from newer tech nodes

Application-system co-design use application-specific features to optimize the BCI system

Form factors **Secondary power effects and thermal constraints** **Wireless power transfer**

GitHub Repo

Paper

Questions?

Thank you!

guyeichler@cs.columbia.edu

https://github.com/GuyEichler/MINDFUL_MICRO25