Heterogeneous Edge for Energy Efficient Computing

… and low latency

Ryan N Coffee / Sr. Research Scientist / LCLS-PULSE-TID

September 11, 2024

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY

Technology (and energy consumption) compounding

A phase transition is coming

CLIMATE

Will A.I. Ruin the Planet or Save the Planet?

It's a notorious energy hog. But artificial intelligence can also foster innovation and discovery, and it could speed the global transition to cleaner power.

By Steve Lohr

CLIMATE

A.I.'s Insatiable Appetite for Energy

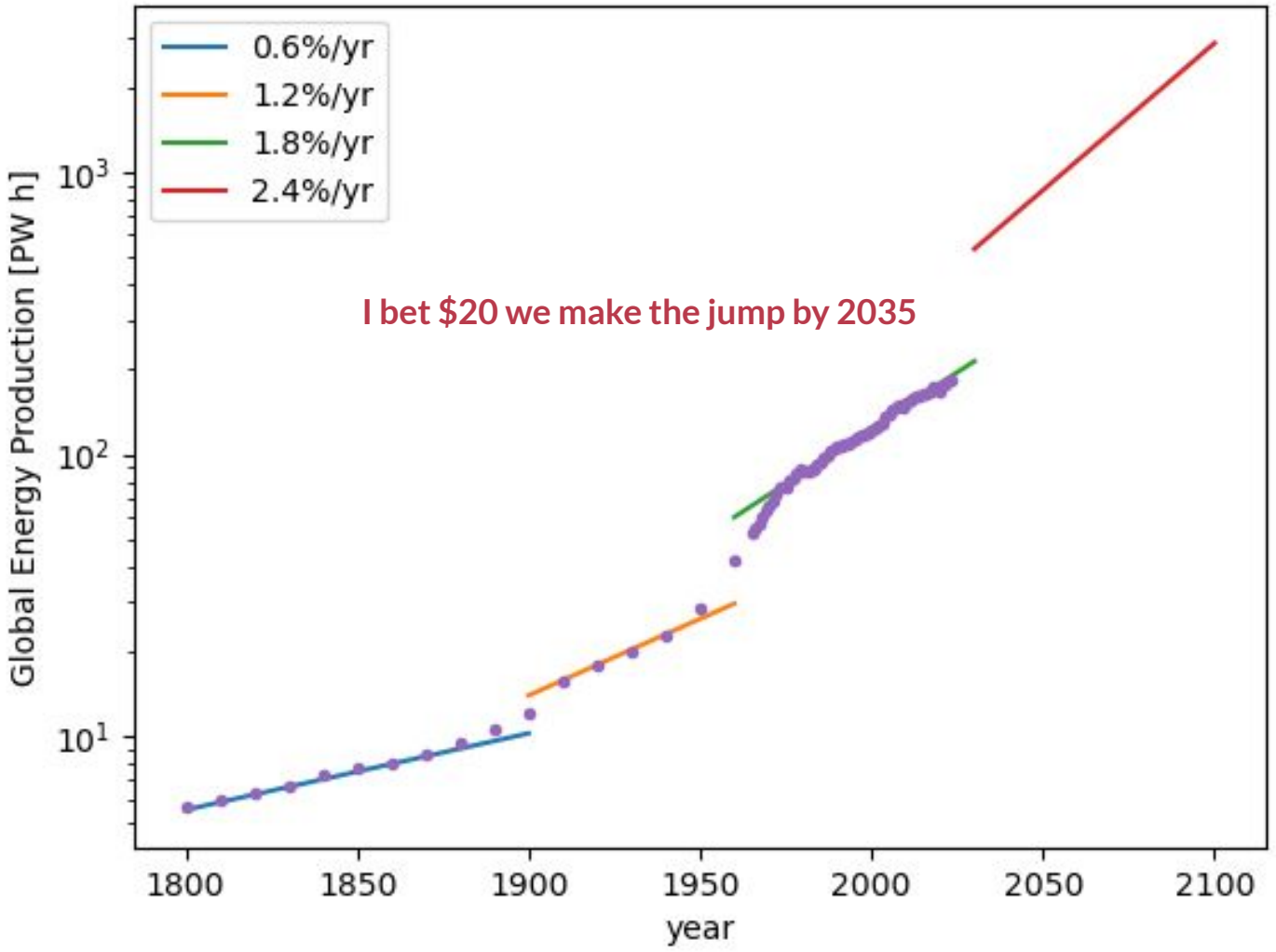
The soaring electricity demands of data centers and A.I. are straining the grid in some areas, pushing up emissions and slowing the energy transition.

By David Gelles

The Climate Summit Embraces A.I., With Reservations

The idea of using artificial intelligence to fight emissions has made a splash at COP28, but there's a catch: The energy it requires could make matters worse.

By Jim Tankersley

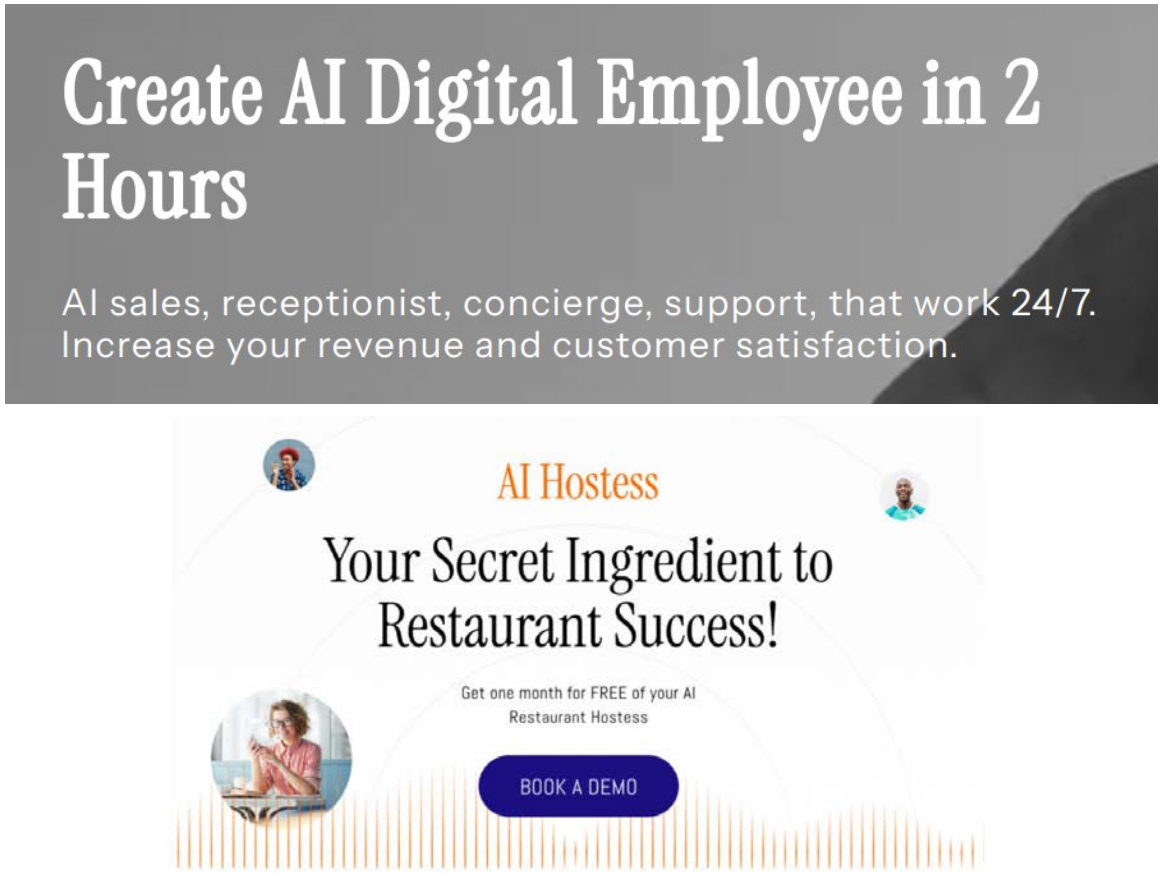


data points: www.nytimes.com/2023/12/03/climate/artificial-intelligence-climate-change.html

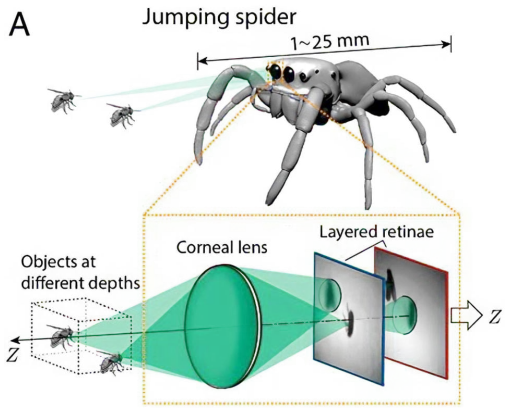
Time is up, Digital Agents are here now!

Invent the future of computing NOW!

- Scientific “Clippy” is inherently **multi-modal**
  - Cross-domain reasoning DNN
  - Scientific tokenizations could “densify” the representations
- Tokenization scheme is far more diverse than with human natural languages
  - How many **alphabets** do we have?
  - How many **sensor signals** do we have?
  - How many **vector-spaces of functions** (with which to represent those signals) do we have?
- **Centralized scientific AGI is fraught,**
  but Edge-to-Exascale is not!



The Parsimonious Jumping Spider (100k neurons)



Eons of co-design

- Hardware and wetware work in unison
- Retinal cells ARE neurons, so are base of each hair on her body, they are acoustic sensors

- Not just computationally efficient… **energy efficient by minimizing bit flow**
- Only outliers are promoted (in humans) to prefrontal cortex (and late)
  - Why **waste** so much computation only for **rationalization**

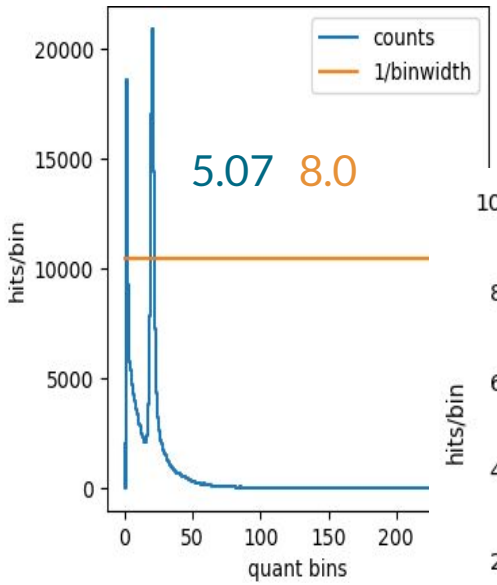
GenAI aims (and misses) reasoning

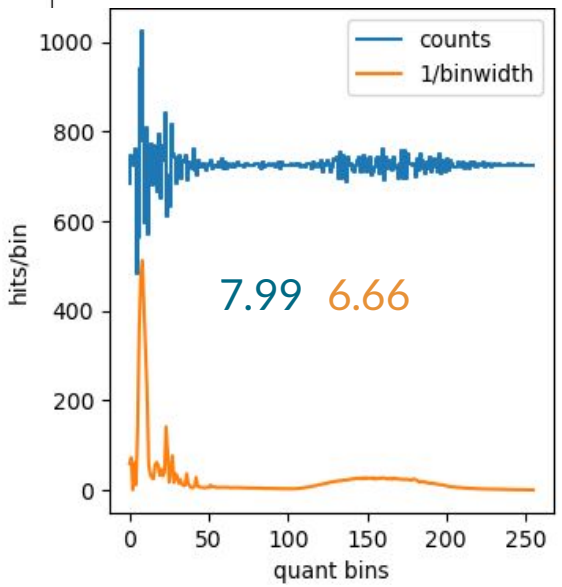
- Aims to learn interpolative “logic”
- But our critical use cases need a formula one pit crew
  - **Performance not Rationalization**

Quantization at the sensor

Rotate Information to Static Metadata

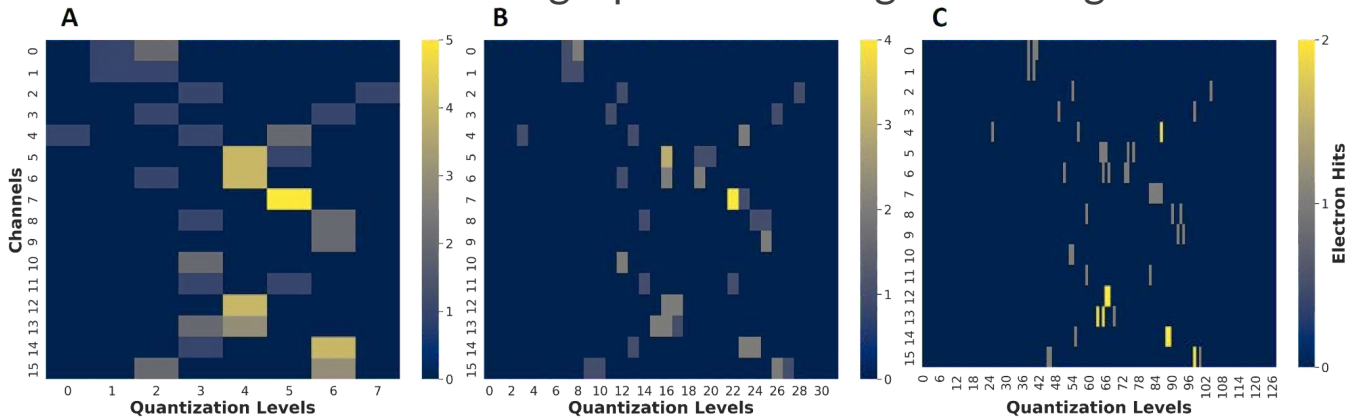
- Prior distribution informs binning
- FPGA (ASIC/Analog?) enforces binning
- Stochasticity of output spectrum is a metric of “concept drift”

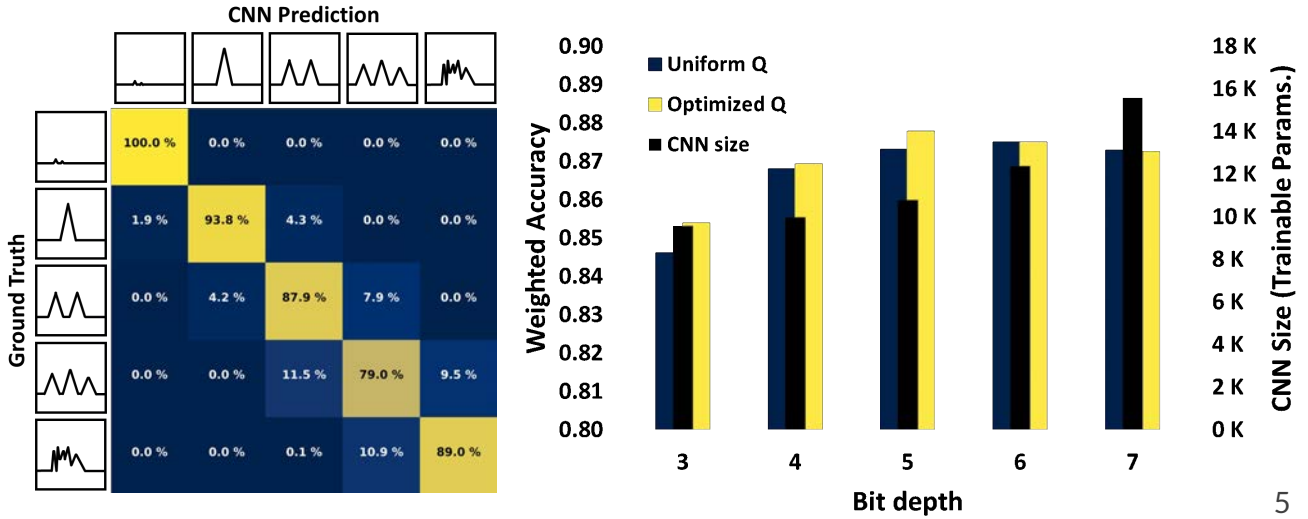




Maximize information/bit

- Far fewer, information dense, input features
- Dense LinAlg Ops for encoding/tokenizing





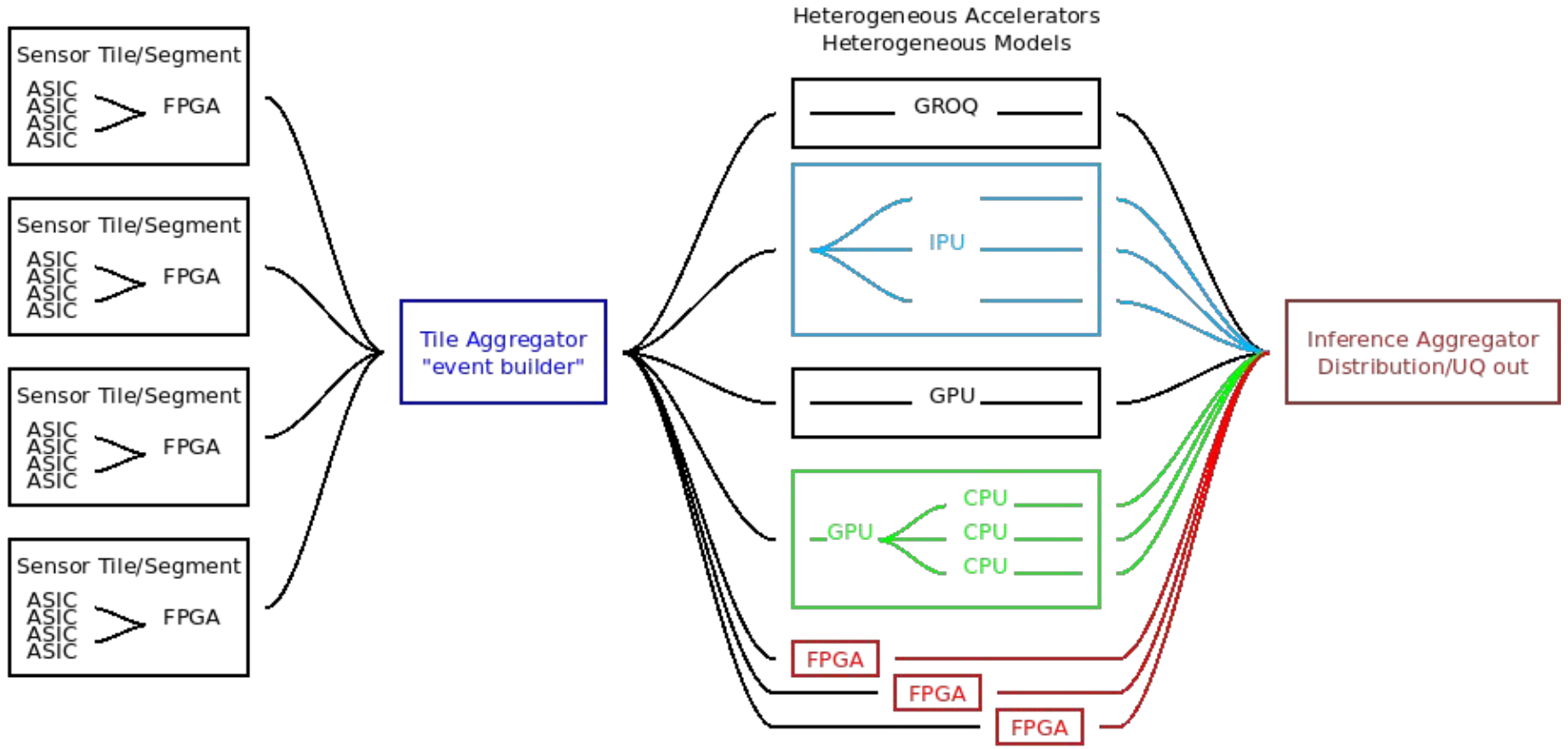
Gouin-Ferland, Coffee and Therrien, Front. Phys. 10 (2022)

Heterogeneous Everything

Orthogonal models are like orthogonal minds

- Each architecture supports a different algorithm module
- Composability of modules allows flexibility
- Orchestration based on hardware simulators and then on real-time module metrics
- ASICs + FPGAs at the sensor edge

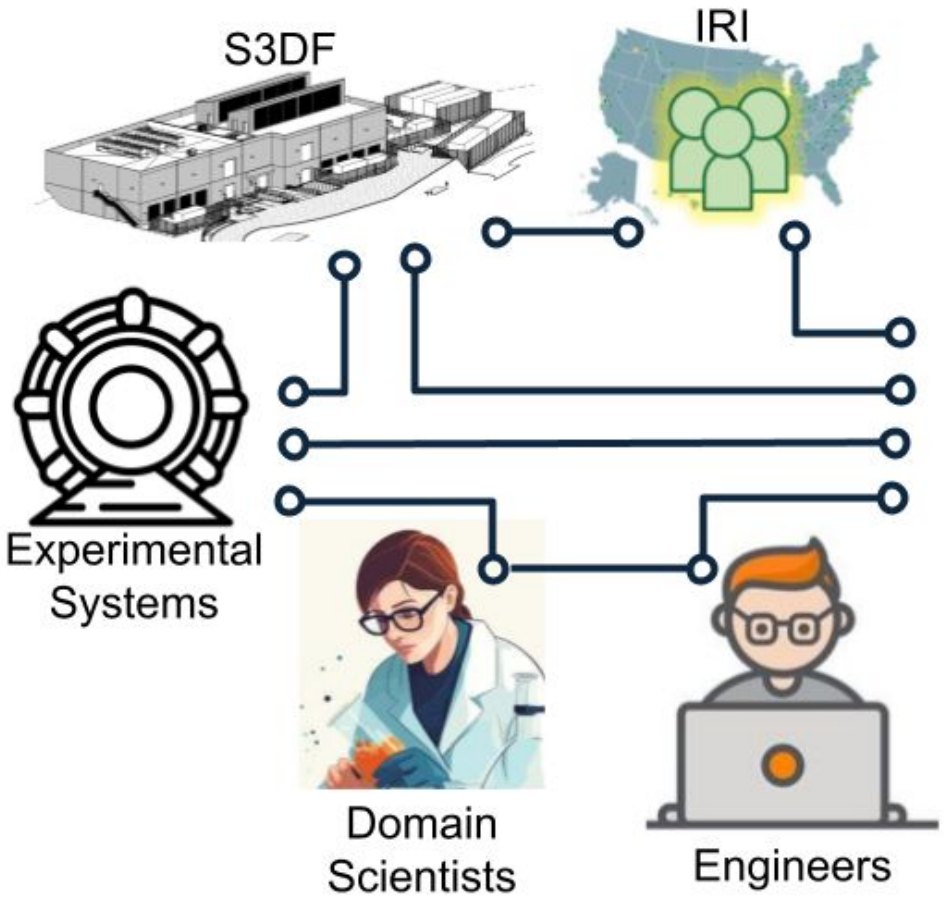
  … or something else?

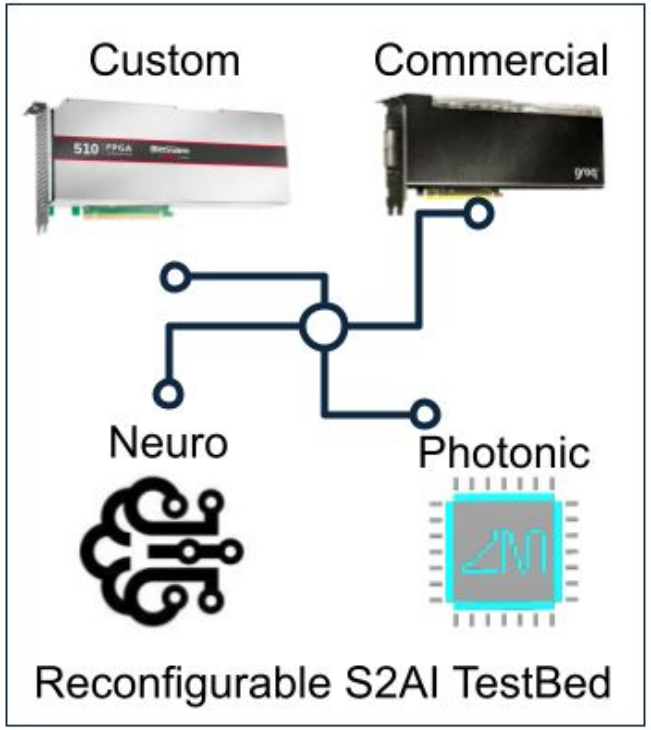


Spectrum from Edge-to-LCF and back!

Domain Scientists and HPC and ASIC Engineers and Researchers

- Tiered Facilities
  - Experimental sensors
  - Mid-scale HPC – also archival storage
  - LCF
- Community Collaboration
  - Workforce Development
  - Open the hood on weird hardware
  - Humanity saving mission



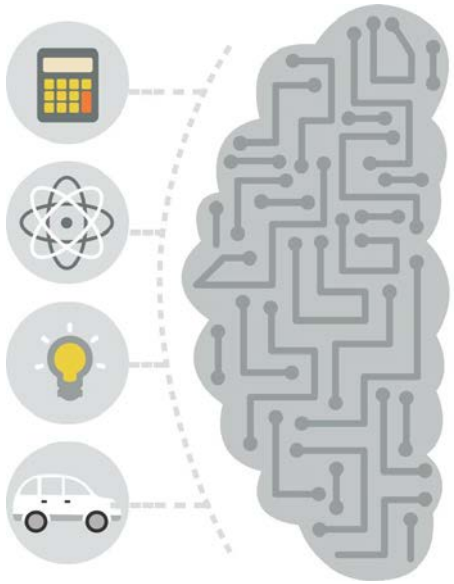


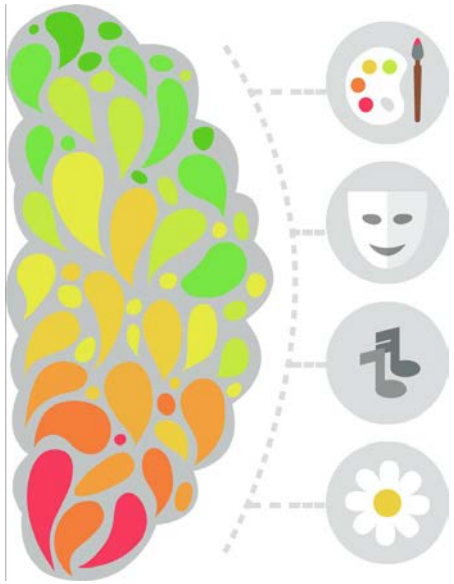
Spectrum from Edge-to-LCF and back!

HPC testbeds linked to Edge Streaming Sensors and Early Access Hardware

- HPC Testbeds for future LCF design for Edge Integration
- Real-world streaming tests to work out bugs and security
- Prototyping inter-lab federation now
- Laying ground work for IRI Orchestration POCs

- Reconfigurable hardware and racks
- Streaming imaging and digitizers (ready for breaking)
- Early access for Edge Inference hardware and custom ASICs and HEP/BES sensor prototypes
- Long DOE FPGA history for trigger and control systems



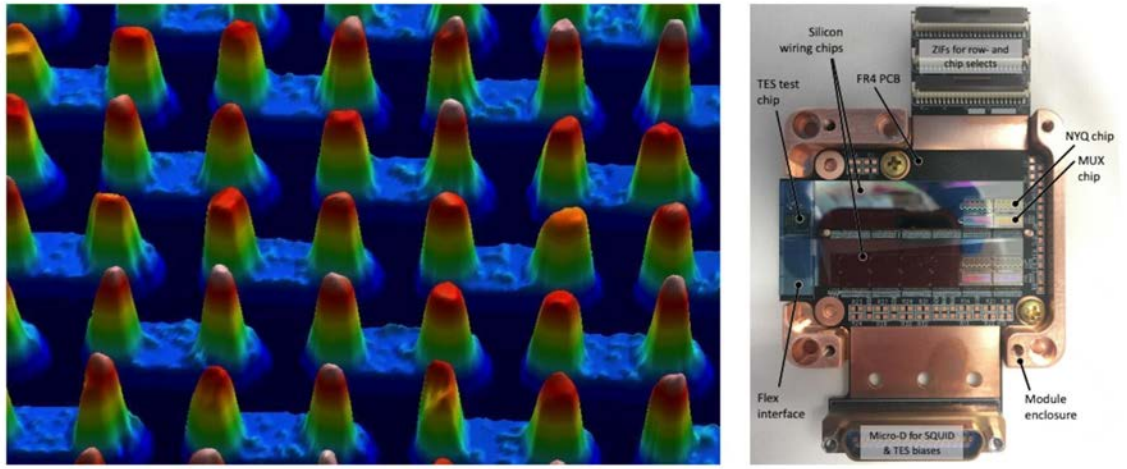


Heterogeneity at the sensors

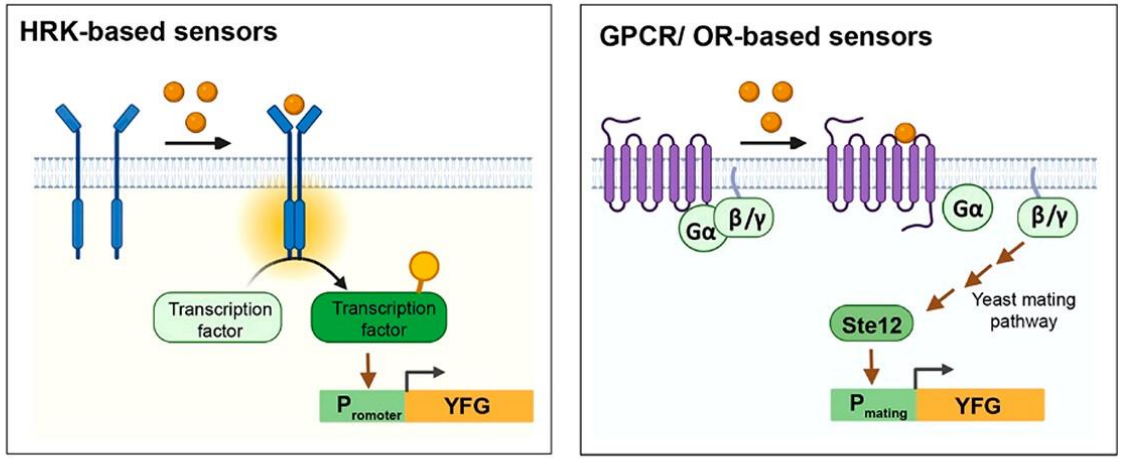
Chiplets / eFPGAs

- Process incoming sensor signals immediately
- Removes the energy burden of moving bits
- Staged control decisions already in the cryo environment,

  … cryogenic magnet controls for tokamak



(a) Indium bumps at 10 um pitch and (b) CMB-S4 prototype assembly. Both fabricated at SLAC. Tomada, A., J. Segal, J. Hasi, C. Kenney, and K. Nishimura. "Flip chip assembly for cryogenics and flexible substrates." In 2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1-3. IEEE, 2015.



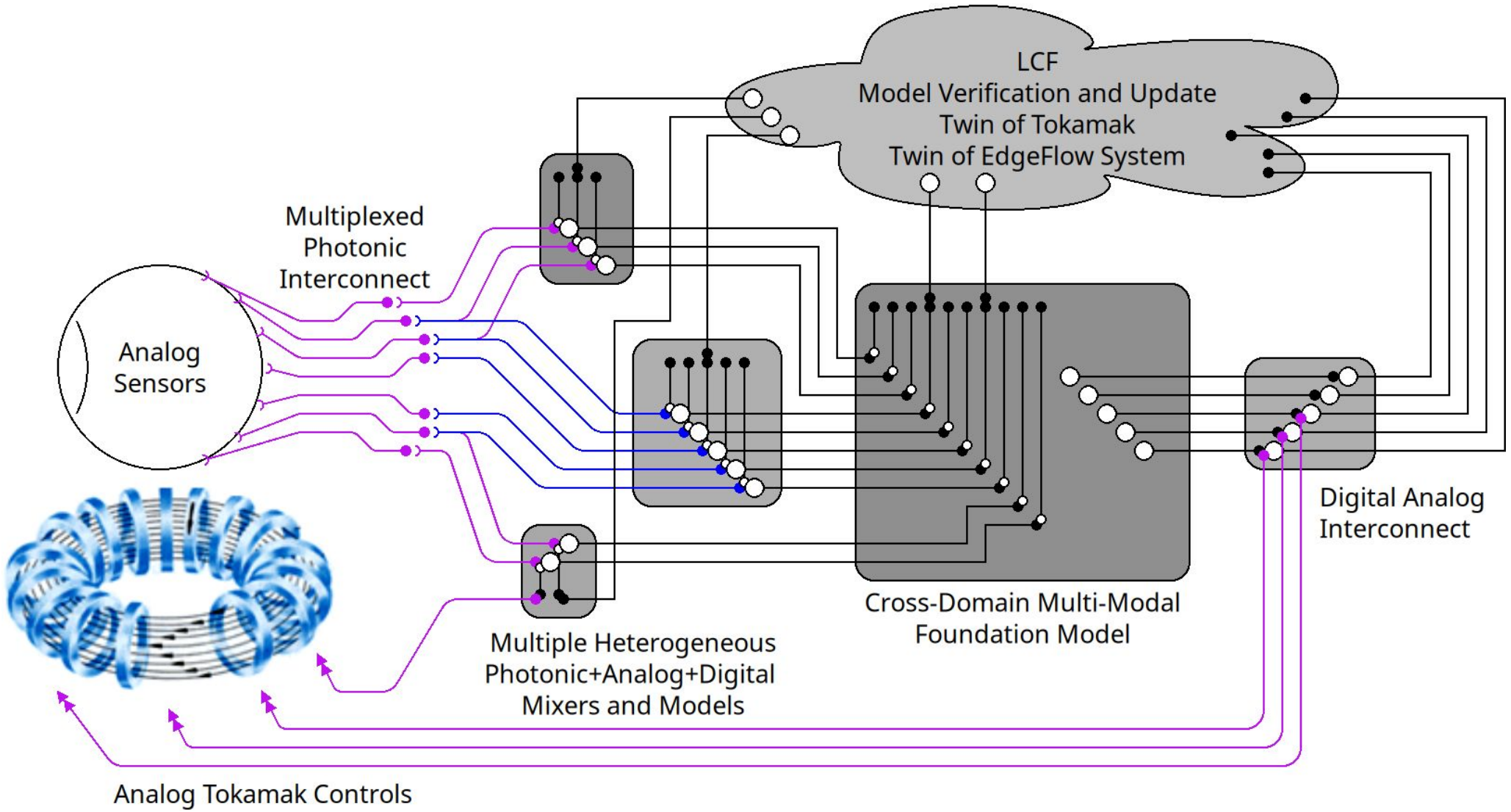
pubs.acs.org/doi/10.1021/acs.biochem.2c00486

Olfactory cells engineered to fluoresce

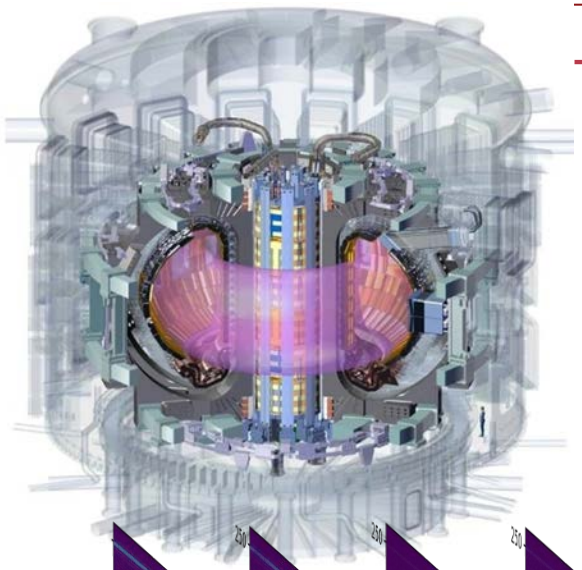
- Photonic receptors and logic for chemical environment sensing
- Biological computing “sensor edge”
- Wild heterogeneity

Biosense-photonic-analog-FPGA-Accelerator-HPC

Spectrum from Edge-to-LCF and back!

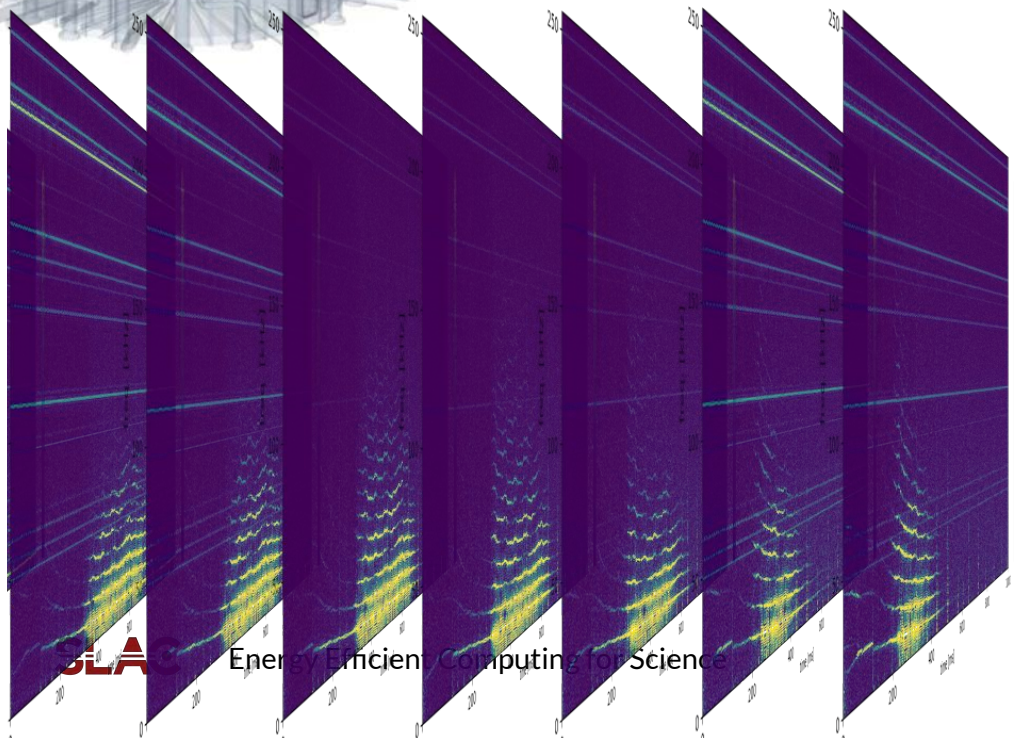


Distributed sensors – Distributed computing



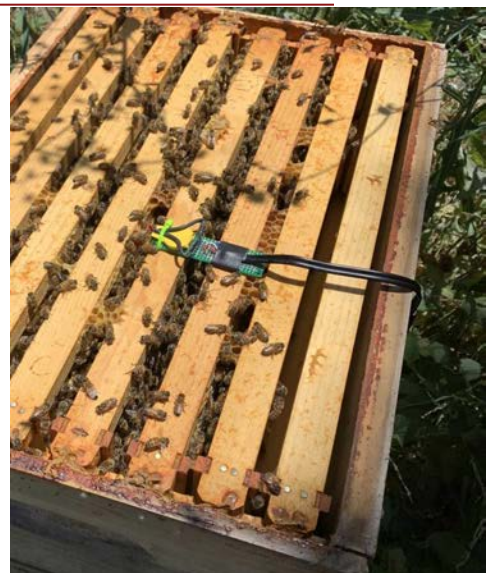
Tokamak magnetics

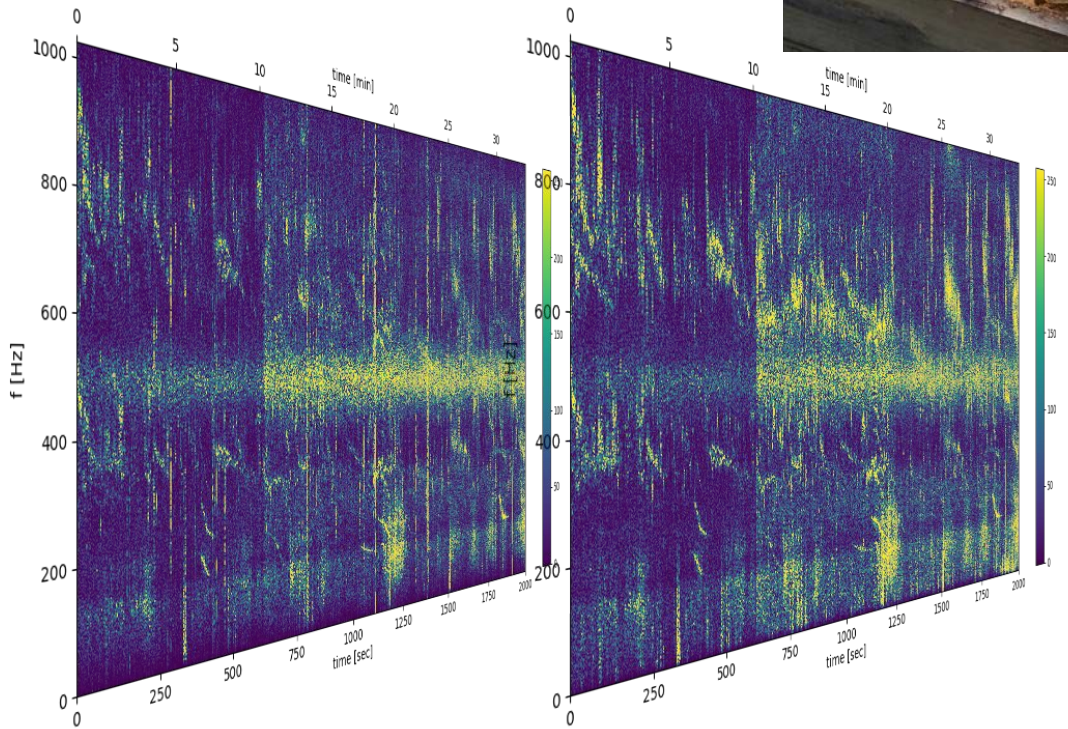
- Disruption **forecasting**
- Need **microsecond latency**
- Real-time controls fed by both **live and local** signal streams and **LCF twins**



Honeybee Acoustics

- **Natural** environmental sensors
- Signals functionally similar across **FES/BES/BER cases**
- ASCR build the tools to pull **all communities** into a Nation Scale computing ecosystem





Nation Scale Computing

Computing could be its own energy cure

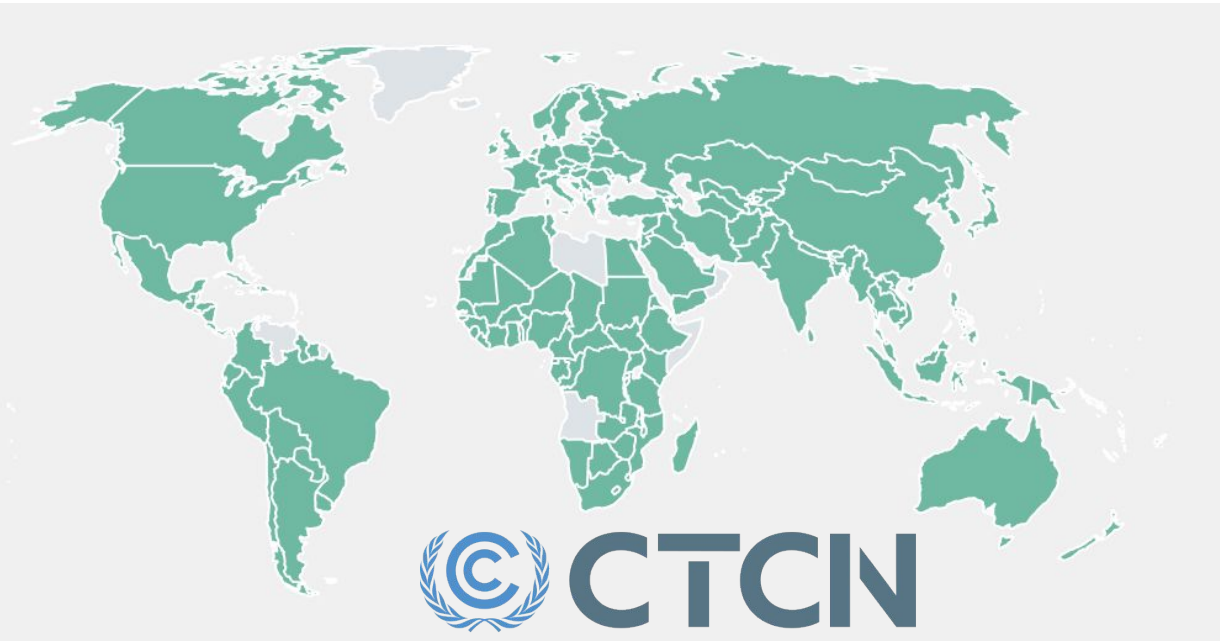
- Infrastructure enabling **fusion forecasting** could enable **climate forecasting**
- Biosensors+biocomputing begs to solve the **federation challenge**

AIM FOR CLIMATE

The globe is watching (and racing) us

- We leverage **computing diversity** just like societies leverage **cultural diversity** and Nature leverages **biodiversity**

  for resilience and efficiency through **real-time adaptation**



Technology (and energy consumption) compounding

A phase transition is coming

CLIMATE

Will A.I. Ruin the Planet or Save the Planet?

It's a notorious energy hog. But artificial intelligence can also foster innovation and discovery, and it could speed the global transition to cleaner power.

By Steve Lohr

CLIMATE

A.I.'s Insatiable Appetite for Energy

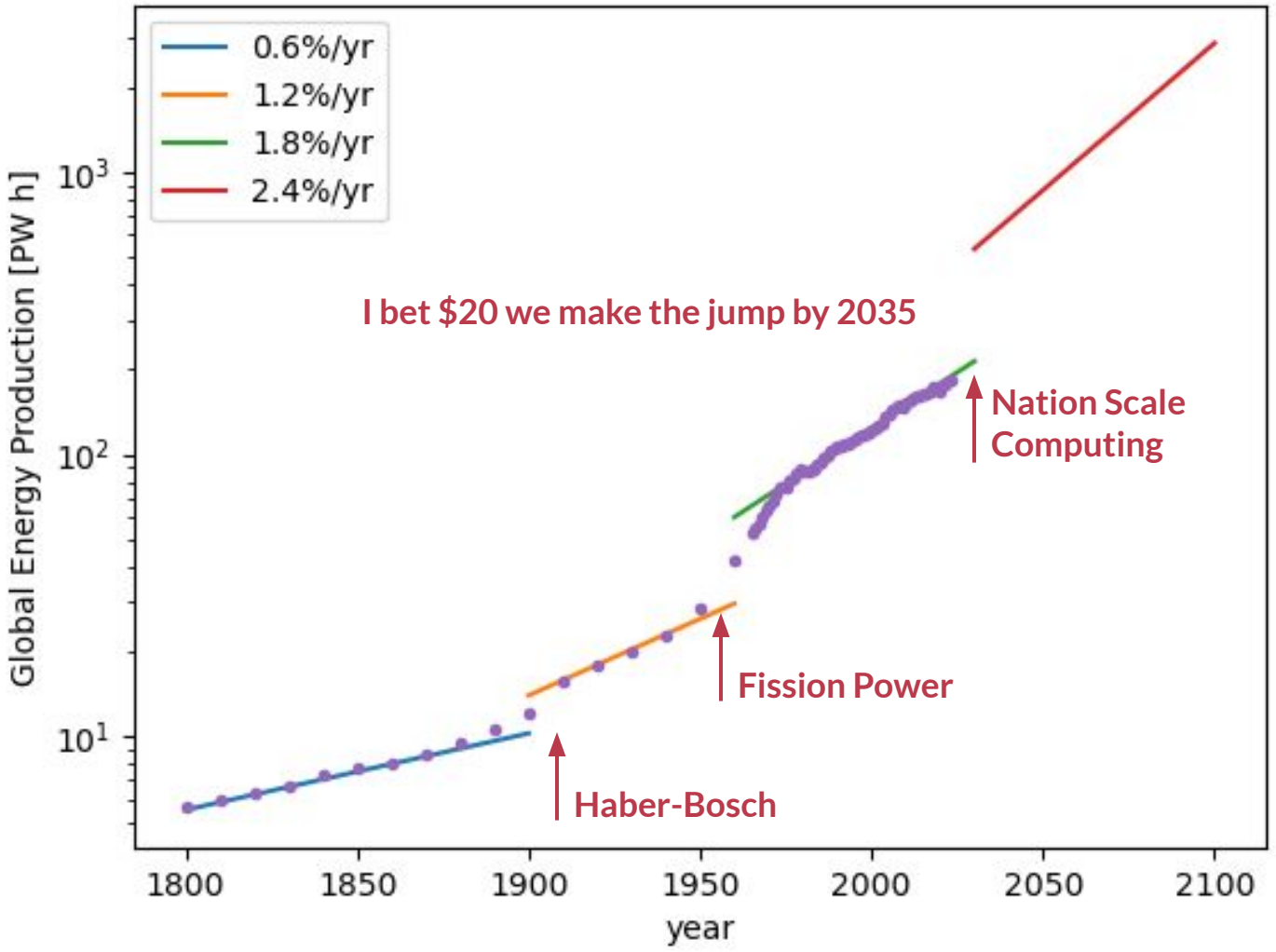
The soaring electricity demands of data centers and A.I. are straining the grid in some areas, pushing up emissions and slowing the energy transition.

By David Gelles

The Climate Summit Embraces A.I., With Reservations

The idea of using artificial intelligence to fight emissions has made a splash at COP28, but there's a catch: The energy it requires could make matters worse.

By Jim Tankersley



data points: www.nytimes.com/2023/12/03/climate/artificial-intelligence-climate-change.html

PRD – Nation Scale Heterogeneous Computing Ecosystem

Opportunity and Direction

- **Inter-lab effort** for real-time Edge-HPC with early access and custom **streaming hardware**
- There is a **global race for computing** dominance, no time to wait
- A Nation Scale computing revolution led by DOE with global impact and demonstration would **close the book on US Technological Leadership**

Execution and Timeline

- Support Edge+HPC **linked testbeds** with **crisp use cases** as benchmark tests
- 5 years: **The IRI Octopus**
- 10 years: **Orchestration** of Heterogeneous Edge processing flow informed/constrained by HPC twins
- 15 years: **Fusion**, Climate Adaptive Agriculture

State of the Art and Challenges

- **Block data movement** to HPC is current tactic
- Edge processing relegated to **isolated test stands**
  - repetition of effort
  - no economy of scale
- **Challenge**: Funding of Edge falls under FES/BES/BER while for HPC it is ASCR

Potential Impact

- Integrated power grid with **Nation Scale Computing**
- Computing infrastructure **as ubiquitous and essential as the interstate highway system**
- Measure of success is **every community** in the US, from inner city to native lands, are using **Edge-to-HPC for their small businesses and agriculture decisions.**

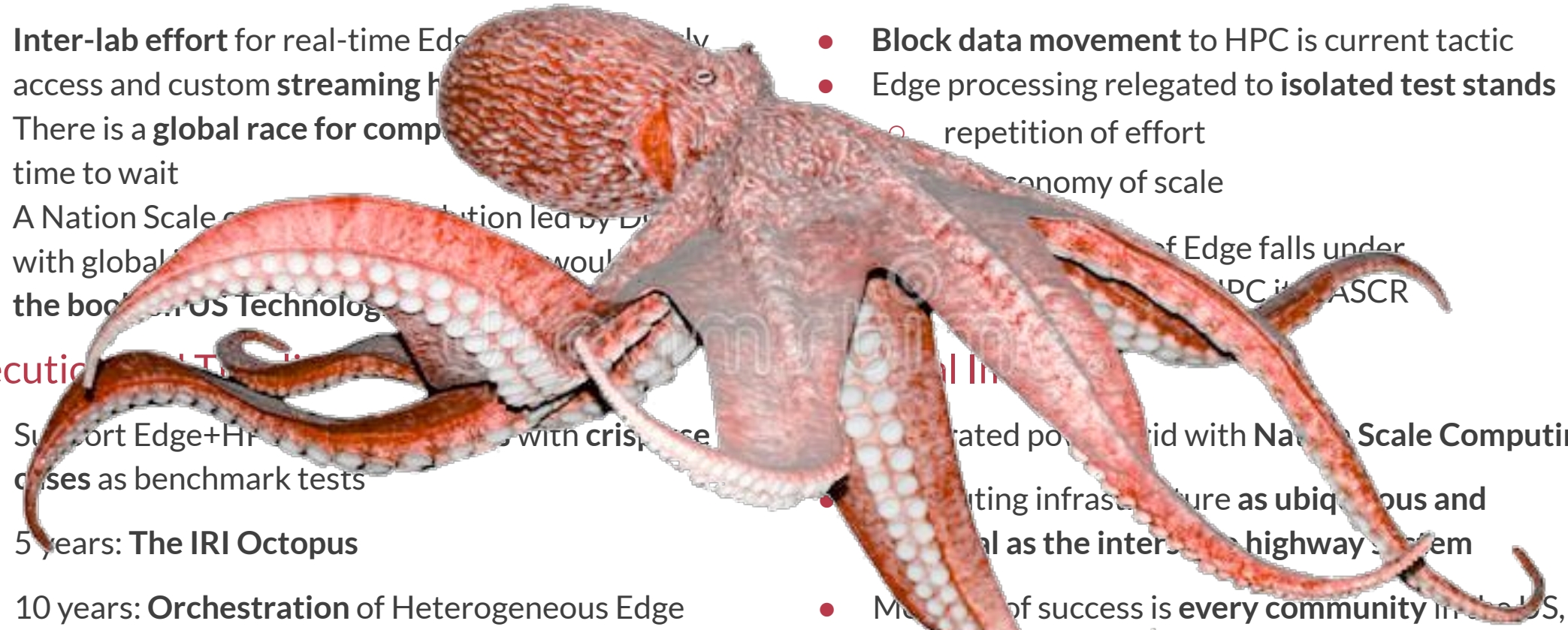
PRD – Nation Scale Heterogeneous Computing Ecosystem

Opportunity and Direction

- **Inter-lab effort** for real-time Ed… …ly access and custom **streaming h…**
- There is a **global race for comp…** time to wait
- A Nation Scale … …ution led by D… with global… …woul… **the boo… OS Technolog…**

Executi… …l Ti…

- Su…ort Edge+Hr… …with **crisp… …ses** as benchmark tests
- 5 years: **The IRI Octopus**
- 10 years: **Orchestration** of Heterogeneous Edge processing flow informed/constrained by HPC twins
- 15 years: **Fusion**, Climate Adaptive Agriculture



State of the Art and Challenges

- **Block data movement** to HPC is current tactic
- Edge processing relegated to **isolated test stands**
  - repetition of effort
  - …conomy of scale
- …f Edge falls under …PC i… ASCR

…l In…

- …rated po… …id with **Na… Scale Computing**
- …uting infras… …ure **as ubiq… …ous and …l as the inter… highway s… …em**
- M… …f success is **every community** in… …S, from i… …city to …ative lands, are using **Edge-to-Hr… …or their small businesses and agriculture decisions.**

PRD – Nation Scale Heterogeneous Computing Ecosystem

Opportunity and Direction

- **Inter-lab effort** for real-time Eds… …ly access and custom **streaming h…**
- There is a **global race for comp…** time to wait
- A Nation Scale … …ution led by D… with global… …woul… **the boo… US Technolog…**

Executi… …l Ti…

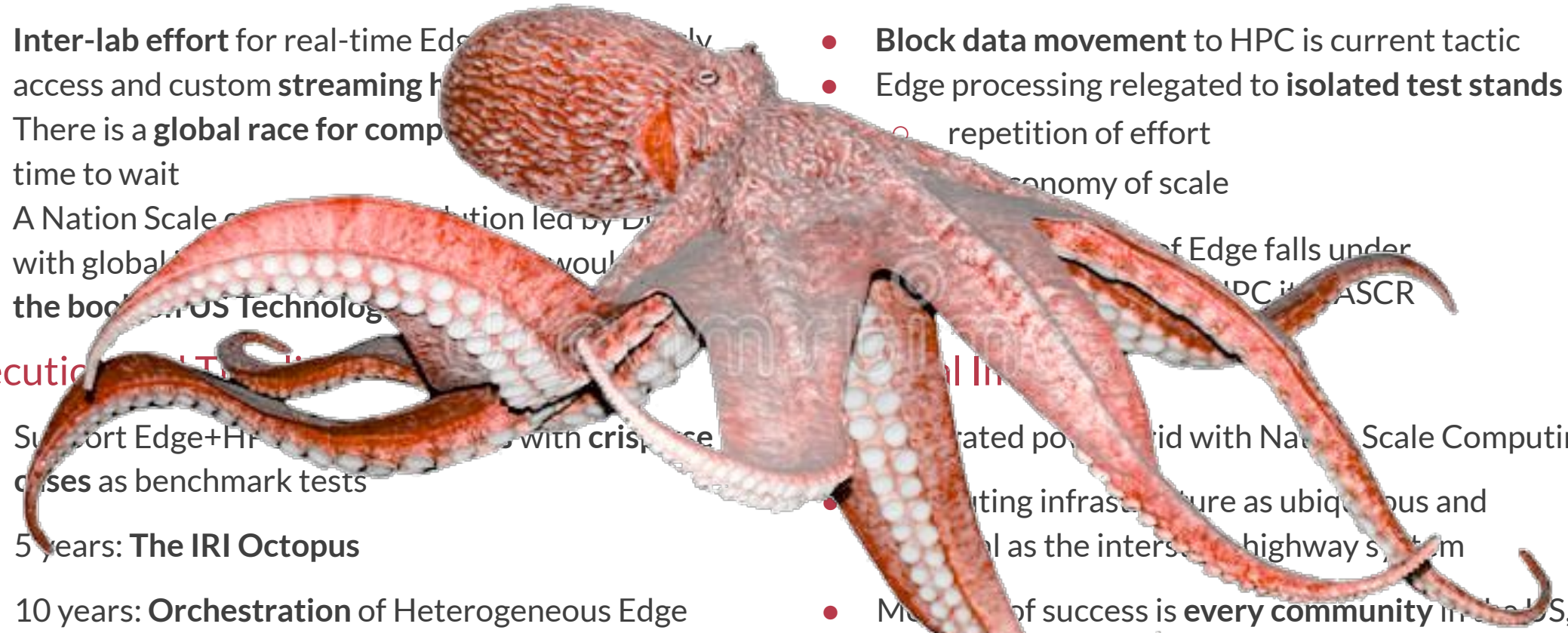
- Su…ort Edge+Hr… …with **crisp… c…ses** as benchmark tests
- 5 years: **The IRI Octopus**
- 10 years: **Orchestration** of Heterogeneous Edge processing flow informed/constrained by HPC twins
- 15 years: **Fusion**, Climate Adaptive Agriculture

State of the Art and Challenges

- **Block data movement** to HPC is current tactic
- Edge processing relegated to **isolated test stands**
  - repetition of effort
  - …conomy of scale
- …f Edge falls under …PC i… ASCR

…l In…

- …rated po… …id with Na… Scale Computing
- …uting infras… …ure as ubiq… …ous and …l as the inters… highway s… …m
- M… …f success is **every community** in… …S, from i… …city to …ative lands, are using **Edge-to-Hr…** for their small businesses and agriculture decisions.



Heterogeneous Edge for Energy Efficient Computing

Ryan N Coffee / Sr. Scientist / LCLS-PULSE-TID

September 11, 2024

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY