Genesis Mission - *Recentering Microelectronics in America*

Lead PI name and e-mail: Gian Michele Innocenti ginnocen@mit.edu

Focus Area: **9F, Microelectronics in harsh environments (High-Energy Physics)**

Tentative proposal title: **Radiation-hard real-time AI processing in 3D-stacked multi-layer tracking devices**

Phase 1 or Phase 2? **Phase 1**

List MIT PI's involved in this proposal:

- Gian Michele Innocenti, ginnocen@mit.edu
- Karl Berggren berggren@mit.edu
- Svetlana Boriskina, sborisk@mit.edu
- Carl V. Thompson, cthomp@mit.edu
- Duane Boning, boning@mtl.mit.edu
- Phil Harris, pharris@mit.edu

Other members of the MIT team:

- Pedro Vicente Leitao (lead digital designer for MOSAIX project—Stitched bendable MAPS for the Electron-Ion Collider @BNL)
- Jelena Lalic (digital designer for MOSAIX project at the Electron-Ion Collider @BNL)
- Ivan Amos Cali, coordinator of the testing & characterization working group of the SVT tracker for the EIC ePIC experiment)

List of collaborators from National Lab, federally funded research centers, industry:

- Grzegorz Deptuch (BNL), Scientist, Group Leader ASIC. Instrumentation Division
- Joao De Melo (BNL), coordinator of the sensor design working group of the tracker of the EIC ePIC experiment
- Kevin Ryu (Lincoln Lab), Assistant leader in the Advanced Imager Technology Group
- Piotr May (BNL) – Instrumentation Staff

Short description of the project

Microelectronics that can operate reliably in harsh environments is a major technological challenge for a broad range of applications including high-energy physics, space instrumentation, defense, nuclear monitoring, and medical imaging. For such applications, future detectors and, more generally, **microelectronic systems must function under extreme radiation, high occupancies, and very large data rates, while remaining low power and low mass and often operating in locations that are difficult, if not impossible, to access**. In this context, High-Energy Physics (HEP) provides a unique benchmark, use case, and proving ground for the development of advanced microelectronics, as it brings together some of the most demanding requirements in radiation tolerance, data throughput, and system integration.

This project will deliver the first realistic implementation study of a 3D-stacked sensor architecture designed with embedded AI processing for real-time trajectory reconstruction. The proposed device will combine multiple closely coupled sensing layers with an integrated AI-processing tier, enabling local noise suppression, clustering, compression, and early reconstruction of particle trajectories at readout. **The key AI novelty is the use of distributed on-chip inference to combine low-level signals from multiple stacked layers in real time, enabling pattern recognition and data suppression beyond the reach of conventional rule-based processing.** The effort will also develop the supporting simulation and validation framework needed to optimize the 3D-stacked device under realistic constraints in radiation, power, area, latency, and bandwidth.

A central objective of this Phase-1 proposal is to demonstrate AI advantage quantitatively. The on-chip AI approach will be benchmarked against traditional offline reconstruction strategies and more traditional digital data-reduction techniques based on reconstructed cluster information. This comparison will establish what AI improves and provides concrete metrics for future development. If successful, the project will open a new direction for HEP detector design based on self-processing, radiation-tolerant sensors.

Extended description of the proposal:

Microelectronics for HEP as an extreme testing benchmark. High-energy physics detectors operate in **extreme environments that combine high particle densities, very large data rates, intense radiation damage, and strict material-budget constraints**. In ultra-relativistic heavy-ion collisions at the LHC, thousands of charged particles can be produced in a single event, requiring extremely fine-grained pixel sensors to keep the occupancy under control. At the HL-LHC, silicon trackers must cope with local hit rates at the GHz/cm^2 scale, and radiation levels approaching Grad and 10^{16} neutron-equivalent per squared cm. **These systems are embedded in large, nearly hermetic detectors installed tens of meters underground, where access is limited and interventions are rare**. After prolonged irradiation, detector components and nearby materials may also become activated, making maintenance more difficult and costly. At the same time, detectors must remain as light as possible to preserve spatial resolution, which requires very low power dissipation to avoid bulky cooling systems, services, and support structures. The current frontier in this area is represented by ultra-light pixel sensors, which combine high granularity with very low material budget and are emerging as the technology of choice for next-generation vertex and tracking detectors. **In this context, performing in-situ data reduction and compression with ultra-low-power devices is no longer optional, but a fundamental requirement to make**

future large-area tracking detectors technically and scientifically viable. Combining the unique know-how developed in high-energy physics with the most recent advances in microelectronics, on-chip AI, and sensor fabrication and packaging could enable a quantum jump not only in scientific discovery, but also in a broad range of other fields that face similar challenges in low-power sensing, real-time data reduction, radiation tolerance, and operation in harsh or inaccessible environments.

The outstanding data-acquisition demands of the Large Hadron Collider have already demonstrated the unique advantages of using AI on hardware, in particular on FPGAs, for fast, low-latency data processing and compression. High-energy physics is one of the fields that pioneered the use of advanced quantized AI models in online reconstruction, triggering, and data-reduction applications, showing that sophisticated inference can be deployed within strict constraints on latency, power, and bandwidth. A large fraction of modern real-time reconstruction tools in HEP now builds on these concepts, and the methods, software, and hardware-design strategies developed by the field have already been exported to a broad range of applications beyond HEP. The main expected benefits of this approach are threefold:

Overview of the proposal. The essence of this proposal is to realize, for the first time, advanced on-chip AI reconstruction inside a radiation-hard 3D-stacked sensor. While ongoing efforts in the field have so far been limited to processing data from a single sensor layer, this project will take the next step by targeting a vertically integrated multi-layer architecture. The novelty therefore lies not only in the development of a low-power 3D-stacked sensor for radiation-hard applications, but also in the ability to preserve and exploit low-level correlations across multiple closely coupled layers before data are transmitted. This will allow us to quantify and characterize the advantage that on-chip AI provides over conventional local digital processing for cross-layer pattern recognition, adaptive filtering and compression, and early trajectory tagging under realistic constraints in timing, radiation tolerance, power, area, and bandwidth.

- **Enhanced multi-layer data reduction** Current sensor readout architectures are typically limited to local thresholding and zero-suppression applied independently at each sensing layer. By exploiting correlated information across vertically integrated layers, the proposed approach enables on-chip data filtering and early inference, significantly improving data-reduction efficiency beyond independent hit processing. By combining multi-point information, this strategy would enable a much more aggressive data-reduction scheme, since redundant geometry can help mitigate the effects of radiation damage. Redundancy across multiple sensing layers would also help maintain robust performance as individual layers degrade under radiation exposure, while supporting the development of self-calibrating procedures that make the detector far more autonomous and therefore especially attractive for extreme applications such as space, where access is not only difficult but sometimes impossible.

- **Reduced power and bandwidth demand** Data movement is the dominant contributor to power consumption in modern sensing architectures. By performing local processing and data reduction at the point of acquisition, the proposed approach reduces bandwidth requirements and enables shorter integration windows, allowing new trade-offs between power, latency, and system complexity.

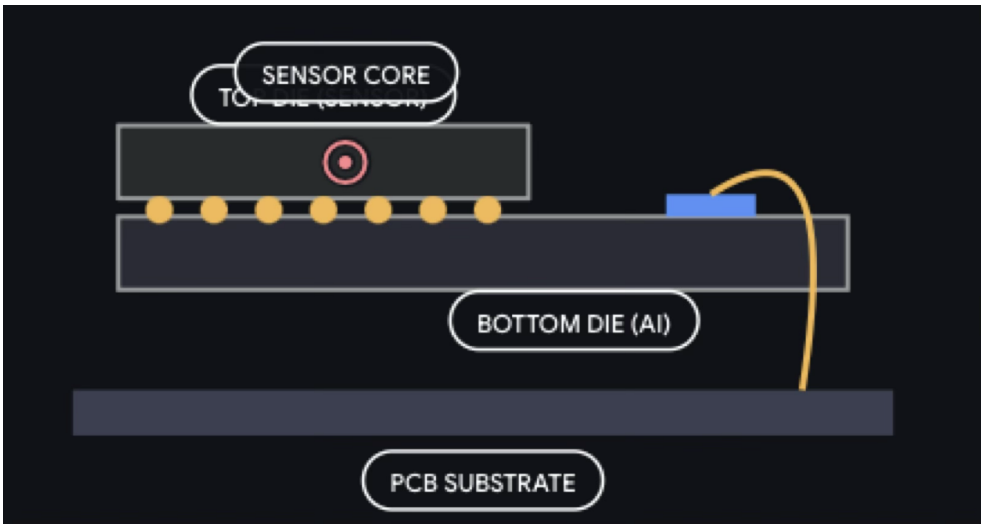
This HEP-driven development will open the way to distributed radiation-hard AI across multiple

integrated layers, increasing the effective computing capability available at the sensor without generating unsustainable local heat dissipation. Although motivated by the needs of collider detectors, the same concept could seed broader advances in sensing and microelectronic systems operating under extreme environmental and system-level constraints.

Methodology and Deliverables:

Design of a 3D-stacked sensor, optimized for operating in a radiation-hard environment under the effect of high magnetic field.

The proposed device is a 3D-stacked multi-layer sensor designed for future tracking detectors operating in harsh radiation environments. Its architecture is based on the vertical integration of sensing and processing tiers through face-to-face bonding using micro-bumps, with the processing tier studied in a 28 nm CMOS technology, enabling a compact and highly interconnected structure. A central requirement of the project is the development of a radiation-hard sensor concept and the corresponding optimization of the stacking strategy so that the integrated device can preserve performance under high-dose, high-hit-rate conditions. The current concept represents a deliberately minimalistic first demonstrator, designed to provide a practical, cost-effective, and scalable entry point into stacked sensor development. In this baseline configuration, in order to reduce development time and costs, the system avoids TSVs in the first implementation by relying on wire-bond connections to the PCB, and targets the emulation of a pixel pitch of about 20 micrometers, a micro-bump pitch of about 40 micrometers, and die dimensions of roughly 5x5 mm2 for the lower tier and 4x5 mm2 for the upper tier. Dr. Kevin Ryu, one of the sensor-imaging experts at MIT Lincoln Laboratory, has already supported the design of this first stacking geometry and confirmed its feasibility using the current capabilities and equipment at Lincoln Laboratory. MIT Lincoln Laboratory and the RLE experts, including Prof. Karl Berggren and Dr. Svetlana Boriskina, will provide expertise in advanced imager technology, device optimization, and integration to refine the demonstrator design, which is expected to evolve toward a more realistic and scalable stacked architecture. This approach provides a credible path toward future heterogeneous, radiation-tolerant stacked sensors with multiple active layers integrated into a very low-mass, fine-granularity detector structure. **By the end of Phase I it will be possible to have the full architecture designed and ready for electronics verification.** The complete design ready for submission of the mentioned ASIC will be part of a Phase II.



This schematic shows the baseline 3D-stacked device: a top sensing die bonded face-to-face through micro-bumps to a bottom AI-processing die. In the first demonstrator, the stacked assembly is connected to the PCB through wire bonds, avoiding TSVs and enabling a simpler, lower-cost prototype.

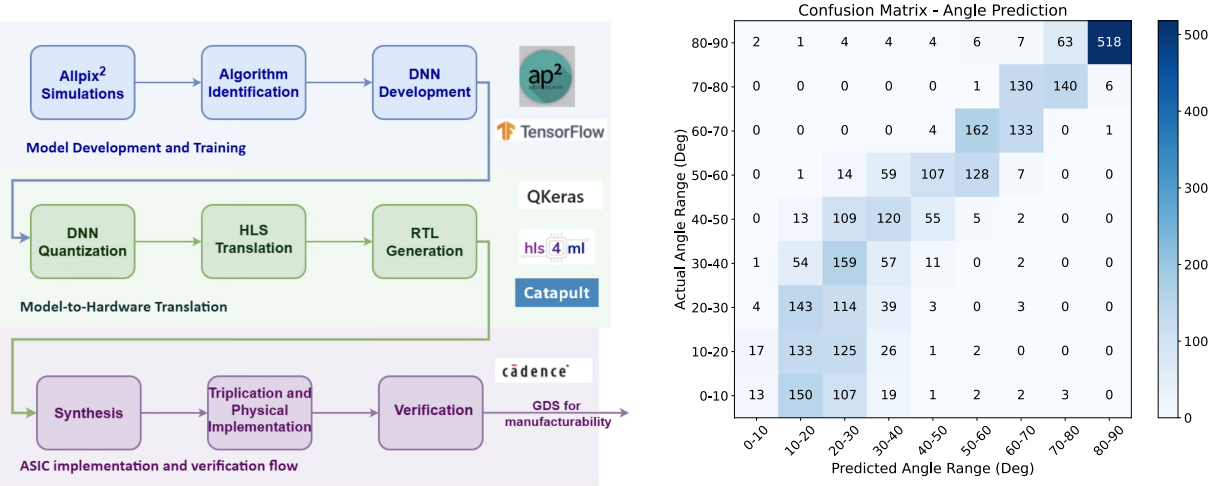
Optimization of radiation-hard distributed AI algorithm to perform on-chip reconstruction of the trajectory of a high-energy particle crossing the sensor.

The AI concept is based on embedding distributed intelligence directly within a multi-layer 3D-stacked sensor so that data from three to four closely coupled MAPS layers can be combined and processed. The targeted functions include noise suppression, real-time clustering, data compression, and early 3D track recognition, with the initial benchmark focused on reconstructing simple particle-trajectory quantities such as the hit origin in the transverse plane and the track angle. The project will also seed the development of general-purpose simulation tools that combine radiation-hard Monte Carlo modeling for stacked devices with the design of an AI-tier architecture, including peripherals, clock management, and realistic hardware constraints. The feasibility of this challenging effort is supported by the experience developed by the MIT and BNL teams in the digital design of what is, to date, one of the most challenging CMOS MAPS sensor developments in HEP, the MOSAIX project, as well as by the experience of the MIT team in designing, optimizing, and translating on-chip single-sensor algorithms into physical implementation and verification. The AI logic will be based on quantized neural network architectures operating on multi-layer sensor data, with an expected complexity in the range of $O(10^3)$ MAC operations, corresponding to an estimated hardware footprint of $O(10^5 - 10^6)$ logic gates[1], depending on quantization and architectural choices. The design will explore low-latency, low-power implementations through reduced precision, time-multiplexing, and distributed processing across stacked tiers.

Simulation and Benchmarking Tools for Radiation-Hard Microelectronics In parallel, the integrated simulation and validation framework developed here will seed the development of a broader set of tools for the design and qualification of radiation-hard microelectronics. By combining radiation-aware Monte Carlo inputs, stacked-device geometry, measured detector-response observables, and realistic hardware constraints from the AI tier, the project will begin to establish a reusable framework for predicting performance, identifying failure modes, and guiding

[1]Estimation performed considering 65 nm CMOS technology

design choices under high-rate and high-dose conditions. This activity will also generate unique benchmark datasets and lay the foundation for an open benchmark data challenge, analogous in spirit to the LHC Olympics effort, for evaluating advanced AI methods in harsh-environment, radiation-tolerant 3D-stacked sensors.



Existing single-sensor AI workflow for 65nm Tower Jazz, showing the full chain from simulation and training to quantization, hardware translation, and physical implementation. This existing framework provides the baseline for future extension to distributed AI in multi-layer 3D-stacked sensors.

Quantitative Evaluation of AI Advantage

A critical component of Phase I will be the definition of a clear non-AI baseline and a quantitative benchmark of AI advantage. The performance of the proposed on-chip AI algorithms will be evaluated against more traditional reduction techniques: first, FPGA-based AI implementations produced through the same end-to-end workflow already developed by the team, and second, standard digital data-reduction techniques based on conventional rule-based or algorithmic logic operating on higher-level reconstructed clusters. This comparison will allow us to isolate what AI improves in practice, namely the ability to exploit low-level multi-layer information for more accurate pattern recognition and stronger data suppression capabilities. In Phase I, AI advantage will therefore be measured quantitatively through side-by-side comparisons of compression factor, reconstruction performance, resource usage, and robustness under realistic high-rate, high-radiation operating conditions, building directly on the team's existing experience in both FPGA-based AI deployment and conventional detector data-reduction strategies.

Additional collaborators

Carl V. Thompson and Duane S. Boning expressed strong interest in joining the proposal and outlined a potential contribution in AI-assisted reliability modeling for harsh-environment microelectronics. In particular, they suggested developing a reliability-analysis module for circuits and systems that uses AI to extract multi-mechanism lifetime models from sparse data, prioritize the most informative failure analyses, and guide the selection of additional accelerated tests needed to qualify materials, processes, and device structures. Their expertise is especially relevant to 3D heterogeneously integrated systems, where multiple concurrent failure modes, including interlayer bond failure, must be isolated and associated with appropriate physical degradation models. They also highlighted the longer-term opportunity to integrate dedicated reliability test structures into stacked devices,

enabling real-time reliability monitoring during operation and the accumulation of datasets for future validation and design optimization. **Their interest is not yet reflected in the current version of the proposal, but this potential contribution will be incorporated in the next revision.**

Outlook and synergistic activities toward a future Phase II effort

This Phase I project is intended to seed a broader program on self-operating, AI-enabled microelectronics for extreme environments. By advancing low-power sensor-adjacent intelligence, stacked architectures, and realistic validation frameworks, it establishes the technical foundation for detector systems that process information locally, reduce data at the source, and remain effective under severe constraints in radiation, data flow, and accessibility.

A natural next step is a common Phase II effort with the parallel Phase I project (targeting 9I) led by Prof. Karl Berggren on materials, devices, and circuits for cryogenic smart-sensor readout. That follow-on activity would explore the extension of stacked sensor and on-chip AI techniques to cryogenic operation, connecting radiation-hard and cryogenic smart-sensor technologies within a unified platform for extreme-environment electronics. Beyond high-energy and nuclear physics, the resulting concepts could impact space instrumentation, cryogenic sensing, nuclear monitoring, medical imaging, fusion and plasma diagnostics, and remote or embedded sensing systems for defense and critical infrastructure. The team is already developing related use cases in Department of Defense projects, and the combined expertise of MIT, BNL, MIT Lincoln Laboratory, and the broader MIT microelectronics community provides a strong ecosystem for translating this work into a wider scientific and technological context.