



THE

Next Wave

Vol. 24 | No. 1 | 2023

The National Security Agency's review of emerging technologies

ISSN 2640-1789 (Print) | ISSN 2640-1797 (Online)



**NEXT-GENERATION HARDWARE FOR
COMMUNICATIONS & SENSORS**

[Photo credit: iStock.com/agsandrew]



GUEST **Editors' column**

Robert J. Runser & David J. Mountain

In this issue of *The Next Wave (TNW)*, we conclude our series reviewing recent advances in NSA's hardware-oriented research. In the last issue, we focused almost exclusively on hardware that will enable the future of high-performance computers, as technologies driven by Moore's Law are sunsetting. In this issue, we review a broad range of novel architecture, hardware, and sensor research that will enable multiple applications including high-performance and secure computing, radio-frequency (RF) monitoring for secure facilities, localizing electric fields for device fault detection, and flexible antenna arrays for detecting multidirectional signals. We also include an article that reviews recent results using additive manufacturing techniques to enable electronics and sensors that adapt and conform to the application geometry and environmental constraints of the system. This issue features authors from multiple research organizations including NSA's research laboratories, the Pacific Northwest National Laboratory, the University of Cambridge, and the University of Maryland. We are extremely grateful for their contributions to this issue.

In the first article, "The road less traveled: Eliminating bottlenecks in high-performance computing networking," the authors provide a historical perspective on the development of different multi-node supercomputing topologies and compare three of the most promising architectures. In their analysis, the authors argue that non-traditional workloads, such as data analytics and artificial intelligence, will require topologies that dynamically remove bottlenecks and adapt to unpredictable workloads driven by new high-performance computing applications.

The next article introduces new extensions to instruction set architectures developed under the CHERI project to address one of the most difficult and long-standing challenges in cybersecurity: memory security. Developed over 10 years, CHERI provides new mechanisms for software developers and hardware designers to enforce fine-grained memory protection to prevent common bugs

that have plagued computing systems for decades. The authors discuss a CHERI prototype for the Arm processor called Morello that will allow evaluation of the new security enhancements and encourage future adoption.

With the end of Moore's Law, new materials and devices will be required to achieve future computing performance gains. Some of these devices will operate at cryogenic temperatures, creating a challenging environment for high-bandwidth interconnects which can dissipate significant heat. The authors of "Evaluating novel interconnects for future cryogenic computers," present their work to establish a test bed for evaluating cryogenic electrical-to-optical devices that provide high-bandwidth data egress from novel devices operating at 4 Kelvin.

Ubiquitous wireless communications protocols and systems have transformed how we communicate. In the article, "Next-generation radio-frequency monitoring in security environments," the authors consider the security risks posed by the wide proliferation of these signals and discuss the RF monitoring requirements to detect and prevent malicious and unintentional emissions that could transmit sensitive data beyond secure facility boundaries.

Localizing faults in today's integrated circuits is essential to improve the manufacturing process but has become extremely challenging due to shrinking feature size and complex fabrication techniques. In "Detecting radio-frequency electric fields with optics," the authors present a novel electro-optic sensor that can detect and localize electric fields with high sensitivity to within less than one millimeter of spatial resolution. These new sensors have potential for a wide range of applications including integrated circuit fault localization and electrical-to-optical conversion of signals.

In "A novel hardware concept for digital beamforming: Development and testing of a frequency multiplexed phased array system," the authors describe a novel

Contents

2 The Road Less Traveled: Eliminating Bottlenecks in High-Performance Computing Networking

SINAN G. AKSOY, ROBERTO GIOIOSA, MARK RAUGAS,
STEPHEN J. YOUNG

10 Improving Security with Hardware Support: CHERI and Arm's Morello

ROBERT N. M. WATSON, PETER SEWELL, WILLIAM MARTIN

22 Evaluating Novel Interconnects for Future Cryogenic Computers

TRISHA CHAKRABORTY, JONATHAN CRIPE, KAREN E. GRUTTER,
GREGORY S. JENKINS, KEVIN D. OSBORN, B. S. PALMER,
PAUL PETRUZZI

34 Next-Generation Radio-Frequency Monitoring in Secure Environments

MINH NGUYEN, BRENT LAIRD, MICHAEL R. GROSS

42 Detecting Radio-Frequency Electric Fields with Optics

KAREN E. GRUTTER, PAUL PETRUZZI, SUMI RADHAKRISHNAN

50 A Novel Hardware Concept for Digital Beamforming: Development and Testing of a Frequency Multiplexed Phased Array (FMFA) System

DAVID ELSAESSER, SPYRO GUMAS, RAVI GOONASEKERAM,
TIMOTHY SLEASMAN, JOHN MARKS

64 Additive Manufacturing of Electronic Circuits for Novel Applications

DANIEL R. HINES

73 Selected Publications by NSA Researchers, 2021–2022

The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government. The views and opinions expressed herein are those of the authors and do not necessarily reflect those of the NSA/CSS.

This publication is available online at <http://www.nsa.gov/thenextwave>. For more information, please contact jsmarx@uwe.nsa.gov.

ISSN 2640-1789 (Print)
ISSN 2640-1797 (Online)

Vol. 24 | No. 1 | 2023



multidirectional RF receiver. This receiver offers the benefits of a conventional phased array system such as high-antenna gain/directivity and co-channel interferer suppression, but does so simultaneously across the entire field of view of the array. This enables detection of weak and short duration signals from any direction. The authors present their design methodology, hardware prototype, and system results from a field test.

In the final article, “Additive manufacturing of electronic circuits for novel applications,” the author discusses new ways to fabricate circuits on nonplanar surfaces to enable electronics and sensors that seamlessly integrate with the geometry of the system and application. The author reviews methods developed at the Laboratory for Physical Sciences for enabling printed electronics and discusses the many benefits including reduced size and weight coupled with the ability to rapidly prototype electronics for field testing.

Advancements in computing and sensing will require flexible architectures, novel hardware devices, and improved sensors that can be integrated together to build systems of the future. The work in this issue illustrates the broad range of research conducted at NSA and with our partners to advance our understanding of these approaches and technologies to enable new mission applications. We thank the authors for their research, which has made this issue of *TNW* possible. We hope you enjoy these articles as much as we enjoyed bringing this issue to print.

Robert J. Runser
Technical Director
Research Directorate, NSA

David J. Mountain
Advanced Computing Systems
Research Directorate, NSA



The Road Less Traveled: Eliminating Bottlenecks in High-Performance Computing Networking

Sinan G. Aksoy, Pacific Northwest National Laboratory (PNNL)


Roberto Gioiosa, PNNL

Mark Raugas, Laboratory for Physical Sciences

Stephen J. Young, PNNL

Scientific, engineering, and social real-life applications are often too large and complex to fit in a single workstation, both in terms of memory and computing requirements. Generally, a cluster of individual compute nodes interconnected by a high-performance network is required to solve such problems at the required scale. Ideally, such a system would function as if it were one huge computer, but in practice, because of the differences in access speed for local and remote resources, a complete new programming paradigm is required. In particular, because the access time difference between local and remote accesses could be in the order of 10–100x, it is paramount to effectively minimize and/or hide the latency of remote communication. Additionally, oftentimes multiple compute nodes need to access data on the same remote node (i.e., many-to-one communication patterns), causing network congestion and slowing down the entire application. As a consequence, one of the significant challenges in the use of modern cluster-based supercomputers is how to efficiently, robustly, and quickly handle the necessary communication between the nodes in the cluster. Both current and next-generation supercomputer designs have highly structured network topologies, such as the low-dimensional torus [1], fat tree [2], or DragonFly [3] topology, to have a straightforward routing scheme while attempting to mitigate the traffic congestion in high-communication applications. In many ways, these topologies have evolved and changed in lockstep with the message passing interface (MPI), the dominant programming model for distributed memory supercomputers, and have become tailored for particular classes of problems (i.e., numerical linear algebra and partial differential equations). However, even with modern high-performance network topologies, communication delays are often a significant bottleneck and dominate the overall computation time.

[Photo credit: iStock.com/carterdayne]

An aerial photograph of a city, likely Seattle, is shown with a dark, textured overlay. The image is covered with a pattern of binary code (0s and 1s) in a light blue/grey color, which is semi-transparent and appears to be floating or layered over the city. The city's buildings, streets, and green spaces are visible through the overlay. The overall aesthetic is high-tech and digital.

As a result of the interaction between the structure of internode communication in various classes of algorithms and the underlying network topologies, certain supercomputers gain a reputation for being more or less suited to a certain class of problems. Specifically, most state-of-the-art supercomputers have been optimized for traditional Linpack-style MPI applications which exchange large messages in highly structured (and often localized) patterns. However, as new problems have emerged that require high-performance computing (HPC) resources, for example, large-scale graph analytics and the training of machine learning models, being able to maintain performance on a more varied collection of communication paradigms has gained in importance. This is especially important to consider when executing on large HPC clusters is the only feasible option for modern graph analytics and machine learning workloads that show computation and memory requirements far beyond those available in a single workstation or small cluster. Of particular relevance to graph analytics and machine learning workloads is the communication performance of HPC systems when sending a large number of small, unstructured, and unpredictable messages. Furthermore, for many of these workloads, the communication patterns are only known at runtime as the computation evolves, making it impossible to predict and mitigate network congestion through smart data layout.

Rush hour and computing

The challenges faced by the HPC community can be understood, by way of analogy, through the evolution of urban transportation traffic. Consider, for example, the Seattle, Washington area. Seattle's arterial road networks, such as the I-5 and I-90 freeways, were developed during the "Boeing Boom." At this time,

the area's largest employers were geographically aligned with the natural traffic pipeline formed by the Puget Sound and Lake Washington. However, as new economic drivers emerged within Seattle, the city has become far more polycentric, with numerous hotspot destinations distributed throughout the region. Seattle's road network now has to contend with a daily influx of traffic from the surrounding Redmond and Bellevue into disparate parts of the city. The resulting traffic patterns are less predictable, less structured, and have (unsurprisingly) led to the development of at least 2,675 documented traffic congestion "hotspots."

While the design of communication architectures for HPC systems doesn't have the geographic limitations of traffic like the greater Seattle area, it is still influenced, much like the traffic network in Seattle, by decades of optimizations for a small class of traffic scenarios. Now that new unstructured traffic scenarios have become more prevalent, the old design paradigms are struggling to provide performance for these new workloads.

Fortunately, rather than having to repeat the decades of effort that went into optimizing HPC systems for MPI-style communications, the HPC communications can take inspiration from an industry that already had to deal with problems of unstructured communication—the telecommunications industry. As early as the 1970's, researchers at Bell Labs and IBM Watson Research Center were thinking about the problem of designing non-blocking switching networks in order to cost-efficiently scale telephone exchanges [4, 5]. Fundamentally, this is a question of how to effectively handle the unpredictable and unstructured telephone communication patterns. Eventually, this line of research coalesced around a single idea as being essential to handling

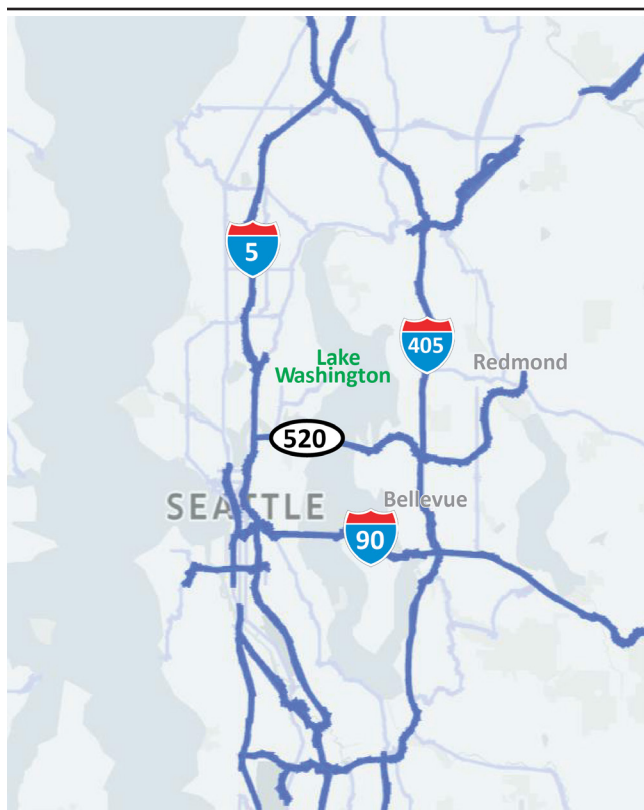


FIGURE 1. This map of the greater Seattle, Washington area road network with the major motorways (no stoplights) highlighted in blue shows lanes in each direction creating a bottleneck. Lake Washington, to the east of downtown Seattle, significantly impacts the topology of the road network, reducing the capacity and number of east-west routes throughout the region. In fact, between the two floating bridges (I-90 and WA-520), there are only five regular traffic lanes and two high-occupancy vehicle (HOV) lanes in each direction.

the unstructured communications of the telephone system—*expansion*. While many definitions of expansion have been proposed over the years, they all essentially reduce to the idea that the capacity of the connections leaving any local neighborhood scale with the size of the neighborhood. Returning to our analogy with Seattle traffic, we can see Lake Washington forms a fundamental obstruction to the expansion of the Seattle road network (see [figure 1](#)). No matter how you increase the capacity of the two floating bridges crossing Lake Washington, or even if you add new bridges crossing the lake, the capacity of the connection from Seattle to the east side will never be able to scale with the size of Seattle. Essentially, Lake Washington forms a geographic *bottleneck* and obstruction to expansion for traffic in

the Seattle area. Surprisingly, this fairly simple idea of considering networks with no bottlenecks has numerous practical applications from constructing circuits to efficiently perform matrix multiplication, to constructing codes which can effectively correct for errors, to methods to amplify weak sources of randomness to high-quality randomness suitable for practical randomized algorithms.

Given the wide applicability of networks with expansion [6], it is unsurprising that several communication topologies have been proposed which use expansion as a fundamental organizing principle. For example, both the Jellyfish [7] and Xpander [8] data-center architectures rely on expansion properties to provide a robust and extensible communication fabric. However, these topologies are fundamentally random in their construction which presents significant challenges in designing and validating the low-overhead communication schemes necessary in computational applications. In addition, the randomness of the connections presents significant obstacles to the adoptions of these topologies in HPC contexts.^a In fact, it is likely that the need for lightweight routing schemes (which are facilitated by highly structured topology) has led to the limited expansion properties of in-use and proposed HPC topologies [9]. However, there are known constructions which result in highly structured, optimal expanders [10]. The SpectralFly [11] topology, which we describe in the following section, is based on one such construction.

The infinite tree in the forest

Before describing the precise construction of the SpectralFly topology, it is helpful to think about exactly what a network with the best possible expansion (or alternatively, no bottlenecks) would look like. Returning to the traffic analogy, imagine traveling on a road network where every intersection is a four-way intersection. As you approach each intersection, you have four choices—turn around and go back along the road you were traveling on, or continue traveling on one of the other three road segments. Now if the road network has the best possible expansion, those three road segments must lead outside your “local neighborhood.” If we imagine continuing along this road network, at each intersection this repeats—you can either turn around or take one of three road segments which leave your “local neighborhood.” But as a consequence, the only way

a. In fact, one of the original proposers of the Jellyfish topology, Brighten Godfrey, obliquely referred to this challenge on his blog *You Infinite Snake*, writing “At this point, one natural reaction is that a completely random network must be the product of a half-deranged intellect, somewhere between ‘perpetual motion machine’ and ‘deep-fried butter on a stick.’”

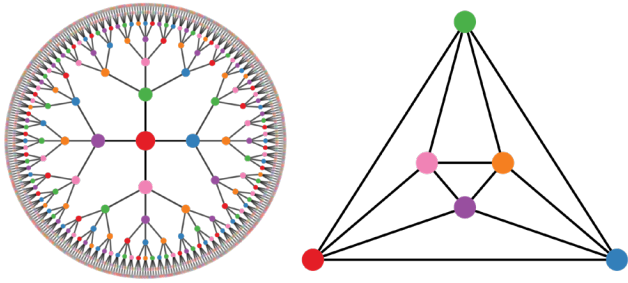


FIGURE 2. (Left) In this optimal expander, every vertex has exactly four connections. Since optimal expanders have no cycles, this unique optimal expander ends up being the four-regular infinite tree. **(Right)** The vertex colors in this vertex-edge graph for the octahedron are used to generate the vertex colors in the infinite tree (on the left) based on the traversals of the octahedral graph.

to return to an intersection you have already visited is to turn around and go back the way you came. In essence, if the road network has the best possible expansion properties, it must be the four-regular infinite tree (see [figure 2](#)).

Obviously, building an infinite tree to use as a road network or as an HPC topology is physically and financially impossible, but taking a slightly different viewpoint on the infinite tree can still provide considerable insight into the properties of networks with good expansion. Specifically, instead of considering the connections in the infinite tree to be physical, we can think about them as a record of decisions made. For example, if we were at the Space Needle in Seattle and wanted to go pick up a coffee at the original Starbucks located at the Pike Place Market, we could either go southwest on Broad Street, turn left on Western Avenue, and continue until we arrived at the Starbucks, or we could go east on Denny Way, turn right on Westlake Ave, and take a right on Stewart Street. While both of these routes will get us some much needed coffee, they emerge from a different sequence of decisions and so would be depicted as different vertices on the infinite tree.

In order to keep track of which locations are the same, we can color individual vertices to encode their location. We see this illustrated in [figure 2](#) where the coloring of the vertices in the infinite tree correspond to the “road network” depicted to the right that has six intersections and 12 roads. For example, in the finite network, the red vertex is adjacent to the green, pink, purple, and blue vertices, and we see that in the infinite tree, every vertex that is colored red is adjacent to a green, pink, purple, and blue vertex. In fact, the correspondence goes deeper than that, as the

colored infinite tree is simply a recording of all the potential routes through the finite network. Indeed, if we start at the red vertex in the finite graph and go to the green vertex, then the pink vertex, and back to the red vertex, in the infinite tree we end up at one of the red vertices in the upper portion of the image of the infinite tree. If, on the other hand, we go to the pink vertex, then the purple, and back to the red vertex, in the infinite tree we end up at one of the red vertices toward the bottom of the infinite tree, despite ending at the same vertex in the finite network. Thus, in many ways, the question of how to design networks with good expansion properties reduces to a perhaps simpler question: *how do you color the vertices of the infinite tree to preserve the expansion properties of the tree?*

To understand what such a coloring looks like, let us consider walking randomly around Seattle. In order to keep track of where we are, imagine every intersection to the west of Lake Washington is colored a different shade of blue, and every intersection to the east of Lake Washington is colored a different shade of red. Since Lake Washington is such a strong bottleneck, it is easy to see that if we start at a blue intersection we should expect to stay on blue intersections for a long period of time. But now think about what this means for the associated colored infinite tree—if we start at a blue vertex, as we go away from that vertex we should typically stay at blue vertices. But there are only so many blue vertices we can use, so that means that the infinite tree must be repeating shades of blue as it grows. In fact, this provides pretty good intuition for comparing two colorings of the infinite tree—a coloring is better at preserving expansion when it is more colorful than another coloring.

Given this framing, it is perhaps not surprising that randomly coloring the vertices of the infinite tree is an effective means of generating graphs with good expansion properties. In fact, this is the approach that is used by the Xpander and Jellyfish topologies to design high-performance data centers. However, this approach has significant drawbacks for HPC needs in that the lack of readily apparent structure in the resulting network means that significant effort needs to be spent in deciding the route any particular communication takes. Providing an explicit means of coloring the vertices of the infinite tree which—in some sense—preserves as much of the expansion property as possible, proved to be a significantly harder challenge.

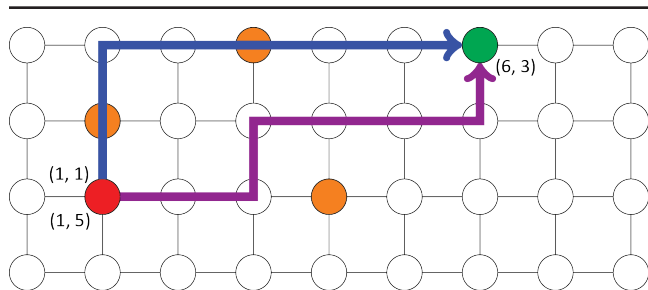


FIGURE 3. This figure depicts an idealization of a downtown street network in a grid-like area such as Manhattan, New York. The blue arrow represents the natural path one would take when going from the red intersection to the green intersection, whereas the purple arrow would be one possible path to avoid construction at the orange vertices.

To understand this challenge, it is helpful to return to the problem of navigating around a road network. However, this time instead of focusing on the freeway system, we will focus on navigating around downtown—perhaps some place like Manhattan, New York, where there is a strong grid-like structure such as shown in [figure 3](#). Imagine your friend calls you from the red intersection looking for directions to your favorite coffee shop, conveniently located on all four corners of the green intersection! How would you tell them to get there? You would probably say something like, head north for two blocks and then go east for five blocks. Or perhaps if you knew they were repairing the sidewalks at the orange intersections, you would tell your friend to go east for two blocks, head north for one block, go east for another three blocks, and finally head north for one more block. Now if the picture in [figure 3](#) was instead a diagram of the switches in an HPC topology and you were providing instruction on how to send information from the red switch to the green switch, you would likely express this idea differently (computers not being particularly well known for knowing which way is north, south, east, or west!). Perhaps you would give each switch a name, say the red switch is switch (1, 1) and the green switch is switch (6, 3), and then you would tell the switches to send the information out the port that increases the second coordinate twice, and the first coordinate five times. In fact, most modern and historical HPC topologies can be thought of in this light. Each switch has a “name”—often a vector of integers—and information is routed by performing a sequence of operations on these names, for example increasing or decreasing a coordinate. Oftentimes, there are additional rules which say two different names are effectively the same. For instance, if we

were to imagine connections between the top and bottom row of vertices in [figure 4](#), we would want to say that (1, 5) is an alternative name for the red vertex, as starting there and increasing the second coordinate four times would return us to the red vertex. Thus, the real challenge is to design a naming scheme for the infinite trees and operations on those names which maximize the colorfulness of the infinite tree.

In the late 1980’s, Lubotzky, Phillips, and Sarnak [[12](#)], and independently Margulis [[13](#)], provided a relatively simple naming scheme and set of operations to provide an optimal coloring scheme for a wide range of infinite trees and number of colors. The collection of names for every vertex is a list of two-by-two matrices with integer entries, and the operation going from one name to another is multiplication by one of a handful of two-by-two matrices. The SpectralFly topology is defined by using one of these networks as the interconnection network between the switches and then placing an appropriate number of compute nodes at each switch (see [figure 4](#)).

The fact that the colorings proposed by Lubotzky, Phillips, and Sarnak [[12](#)] are the best possible at preserving the expansion properties of the infinite tree relies on a deep result in the representation theory of automorphic forms originally conjectured by Ramanujan [[14](#)]. However, we can gain some intuition as to why their rules result in more colorful trees by comparing the operations with other topologies. For example, since the operation for the torus topology is incrementing/decrementing individual coordinates, the end location depends only on the number of increments/decrements per coordinate, not the particular order they are applied. In contrast to this, the results of matrix multiplication (in general) rely on the order of operations. That is, by applying the same set of operations in two different orders, it is possible to arrive in different locations. In [figure 5](#) we can see the difference in colorfulness in the infinite tree for torus topology and the SpectralFly topology.

Structural comparison with SpectralFly topology

The SpectralFly topologies promise as supercomputing topology is evidenced by its exceptional structural properties. We now put these properties in perspective, by comparing them against those of two well-known topologies: a DragonFly network and a

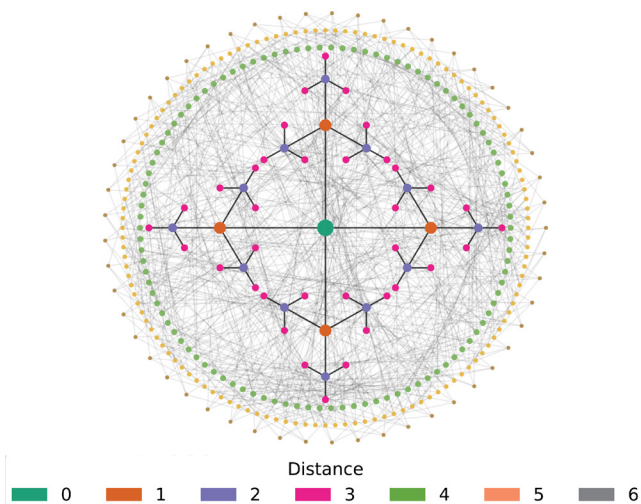


FIGURE 4. In this depiction of the connections between switches for a 336-switch SpectralFly topology with four intra-switch connections per switch, the switches are color coded by distance from a central switch, highlighting the tree-like neighborhood of the switch.

torus mesh. We consider a small, sparse SpectralFly network on 120 nodes and 240 links. To ensure a fair comparison, we optimally select^b the parameters of a DragonFly topology on exactly the same number of nodes and edges, and a two-dimensional torus mesh on 121 nodes and 242 edges. This near-exact three-way match enables a size agnostic comparison: each network starts with the same number of nodes, links, and radix, but makes different design choices in assembly.

The three networks are visualized in [figure 6](#). Each row of [figure 6](#) plots the same network, but with one of three different structural properties emphasized: the tightest bottleneck, the network diameter, and link usage frequencies in random traffic. Each of these structural properties are fundamental for supercomputer design: bottleneckness measures congestion proneness, diameter is a proxy for worst-case latency, and link usage patterns impact link-contention.

Bottlenecks

The first column of [figure 6](#) presents a split of each topology into equal parts which minimizes the number of edges crossing (in red), as found by METIS software [16]. For SpectralFly, Dragonfly, and torus, this yields 40, 31, and 26 links crossing,

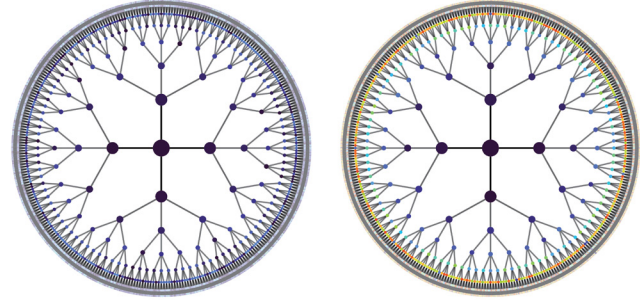


FIGURE 5. In this comparison of an infinite tree for a two-dimensional torus topology (left) and a SpectralFly topology with parameters (3,7) (right), the vertices of both are colored using the same equally spaced gradient. The vertex color corresponds to the order the vertices are discovered in the process; earlier vertices are colored blue and later vertices are colored red. As we can see, the torus topology is significantly less colorful than the SpectralFly topology, indicating that the SpectralFly topology has significantly better expansion properties.

respectively. In practice, this means that when there are many messages, we would expect the communication delays to be about 24% smaller as compared to DragonFly, and 35% smaller as compared to the torus.

Diameter

[Figure 6](#)'s second column visualizes paths linking a source-destination pair furthest from each other in the network—the length of which is known as the network diameter. The k -th ring of vertices from the leftmost contains all those that can be reached from that vertex in k hops. Small diameters ensure any vertex can be reached quickly from any other. In this case, both SpectralFly and DragonFly have an identical diameter of 6, while the torus has a diameter of 10.

Link loading

We simulate unstructured traffic on each network by randomly selecting 5,000 source-destination pairs in each network, and then routing via a minimal path. In the case that there are multiple such minimal paths, we select one at random. For each link in the network, we count the number of times it was traversed. [Figure 6](#) presents the distribution of these link usage counts. For SpectralFly, this distribution is highly symmetric and tightly concentrated, reflecting that edges are evenly spread across the network.

b. In particular, we generate a DragonFly topology with height $h=1$, $g=24$, groups of size $a=4$. We optimally allocate the intergroup edges as suggested by Teh, Wilke, Bergman, and Rumley [15].

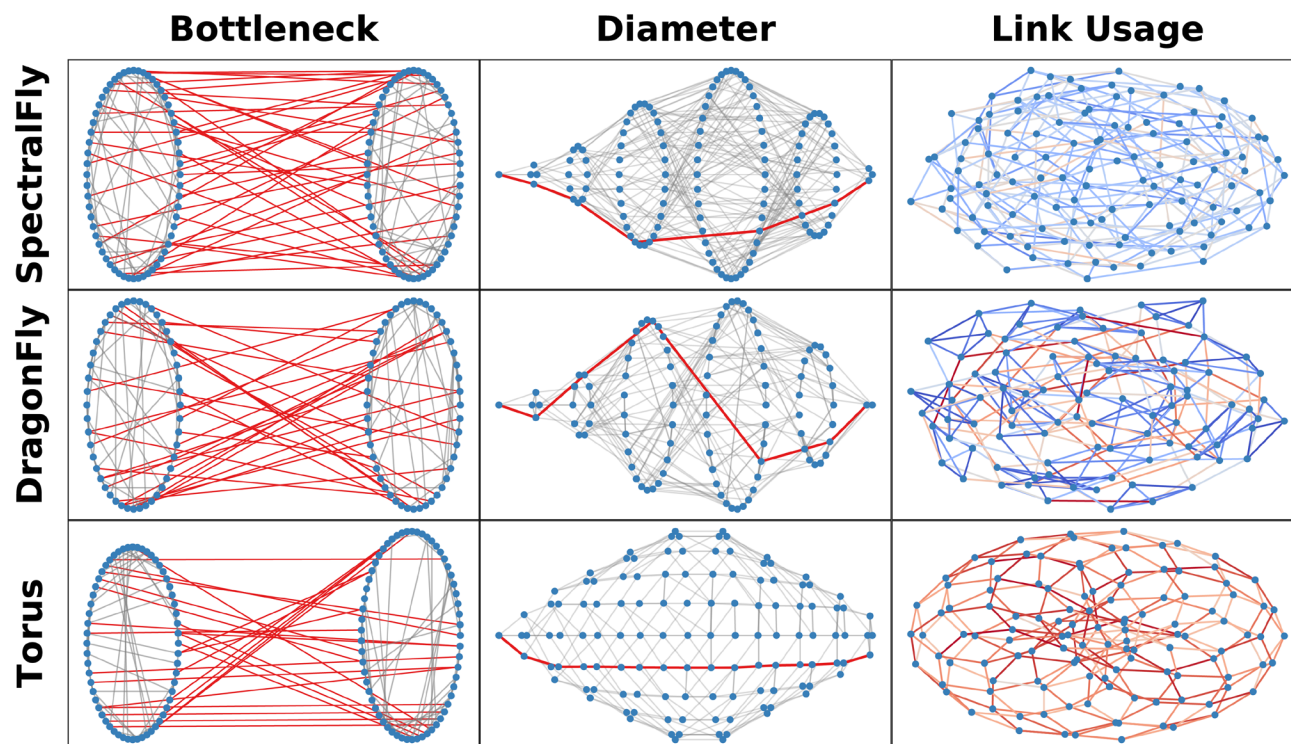


FIGURE 6. These graph visualizations emphasize different structural properties for three similarly-sized SpectralFly, DragonFly, and torus instances, each on about 120 nodes and 240 edges. The first column emphasizes the expansion, the second column emphasizes the diameter, and the third column emphasizes the prevalence of edges on shortest paths.

DragonFly, on the other hand, is the opposite and has a long tail: some edges are used heavily while others are almost never used. Lastly, as with SpectralFly, the link usage counts for the torus are also tightly concentrated, but overall larger: due to the torus' larger diameter, paths linking vertices tend to be longer and edges get used more frequently, albeit evenly, across the network. These observations are also reflected in the network visualization: each edge is colored on a blue-to-red scale, according to its percentile within the observed link counts aggregated across all three networks. Accordingly, SpectralFly and the torus' edges are homogeneous in color; whereas, those in Dragonfly run the gamut.

Conclusions and future work

As the workloads executed on current and future HPC systems evolve to include nontraditional workloads, such as data analytics and artificial intelligence (AI)/machine learning, so should the systems themselves. We argue that HPC systems should move away from

highly structured networks optimized for regular and large message communication to networks such as SpectralFly that expand and dynamically remove bottlenecks, and hence adapt to the irregular and unpredictable nature of the workloads. This move however would be onto a road less traveled, and as such, will require strong evidence that it can efficiently support emerging application domains before industry will commit to investing in it. At Pacific Northwest National Laboratory, we have developed and used several tools, based on MPI and partitioned global address space (PGAS), to analyze different network designs. The results indicate that SpectralFly networks are not only better at supporting irregular communication typical in data analytics and AI/machine learning, but that they might also outperform traditional networks when executing regular applications (unless they heavily rely on near-neighbor communication). In other terms, the SpectralFly network will let you sip your much deserved, end-of-the-day coffee at your favorite coffee shop without spending hours stuck in the car on the streets of Seattle downtown. 🌈

References

- [1] Adiga NR, Blumrich MA, Chen D, Coteus P, Gara A, Giampapa ME, Heidelberger P, Singh S, Steinmacher-Burow BD, Takken T, Tsao M, Vranas P. "Blue Gene/L torus interconnection network." *IBM Journal of Research and Development*. 2005;49(2-3):265–276.
- [2] Leiserson, CE. "Fat-trees: Universal networks for hardware-efficient supercomputing." *IEEE Transactions on Computers*. 1985;C-34(10):892–901. doi: 10.1109/TC.1985.6312192.
- [3] Kim J, Dally WJ, Scott S, Abts D. "Technology-Driven, Highly-Scalable Dragonfly Topology." *SIGARCH Computer Architecture News*. 2008;36(3):77–88. doi: 10.1145/1394608.1382129.
- [4] Chung FRK. "On concentrators, superconcentrators, generalizers, and nonblocking networks." *The Bell System Technical Journal*. 1979;58(8):1765–1777. doi: 10.1002/j.1538-7305.1979.tb02972.x.
- [5] Pippenger N. "Superconcentrators." *SIAM Journal on Computing*. 1977;6(2):298–304. doi: 10.1137/0206022.
- [6] Hoory S, Linial N, Wigderson A. "Expander graphs and their applications." *Bulletin of the American Mathematical Society*. 2006;43(4):439–561. doi: 10.1090/S0273-0979-06-01126-8.
- [7] Singla A, Hong C, Popa L, Godfrey PB. "Jellyfish: Networking data centers randomly." In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*; 2012. Available at: <https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final82.pdf>.
- [8] Valadarsky A, Shahaf G, Dinitz M, Schapira M. "Xpander: Towards optimal-performance datacenters." In: *Proceedings of the 12th International Conference on Emerging Networking EXperiments and Technologies*; 2016, pp. 205–219. doi: 10.1145/2999572.2999580.
- [9] Aksoy SG, Bruillard P, Young SJ, Raugas M. "Ramanujan graphs and the spectral gap of supercomputing topologies." *The Journal of Supercomputing*. 2021;77(2):1177–1213. doi: 10.1007/s11227-020-03291-1.
- [10] Alon N. "Eigenvalues and expanders." *Combinatorica*. 1986;6(2):83–96. doi: 10.1007/BF02579166.
- [11] Young S, Aksoy S, Firoz J, Gioiosa R, Hagge T, Kempton M, Escobedo J, Raugas M. "SpectralFly: Ramanujan graphs as flexible and efficient interconnection networks." In *2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Lyon, France, 2022 pp. 1040–1050. doi: 10.1109/IPDPS53621.2022.00105.
- [12] Lubotzky A, Phillips R, Sarnak P. "Ramanujan graphs." *Combinatorica*. 1988;8(3):261–277. doi: 10.1007/BF02126799.
- [13] Margulis GA. "Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators." *Problemy Peredachi Informatsii*. 1988;24(1):51–60.
- [14] Ramanujan S. "On certain arithmetical functions." *Transactions of the Cambridge Philosophical Society*. 1916;22(9):159–184.
- [15] Teh MY, Wilke JJ, Bergman K, Rumley S. "Design space exploration of the dragonfly topology." In: Kunkel J, Yokota R, Taufer M, Shalf J (Eds), *High Performance Computing. ISC High Performance 2017. Lecture Notes in Computer Science (LNTCS)*, vol 10524. Springer, Cham. Available at: https://doi.org/10.1007/978-3-319-67630-2_5.
- [16] Karypis G, Kumar V. "A fast and high quality multilevel scheme for partitioning irregular graphs." *SIAM Journal on Scientific Computing*. 1998;20(1):359–392. doi: 10.1137/S1064827595287997.

Improving Security with Hardware Support: CHERI and Arm's Morello

Robert N. M. Watson, University of Cambridge
Peter Sewell, University of Cambridge
William Martin, National Security Agency

The CHERI project, from the University of Cambridge and SRI International, extends instruction-set architectures (ISAs) with unforgeable *architectural capabilities*, to be used in place of conventional machine-word addresses to access memory. CHERI, which stands for Capability Hardware Enhanced Reduced Instruction Set Computer (RISC) Instructions, deterministically protects C/C++ pointers and other references, and also enables in-address-space software sandboxing. With changes to the compiler and operating system (OS), CHERI enables new hardware-software security protection models for existing software (typically with only very minor changes for memory safety):

- ▶ Deterministic fine-grained C/C++ memory protection at low overheads; and
- ▶ Scalable software compartmentalization, including sandboxed libraries, with interprocess communication performance improvements and function-call-like domain transition.

In a 2020 blog post evaluating CHERI, the Microsoft Security Response Centre (MSRC) wrote: "We've assessed the theoretical impact of CHERI on all the memory safety vulnerabilities we received in 2019, and concluded that in its current state, and combined with other mitigations, it would have deterministically mitigated at least two thirds of all those issues" [1]. Scalable single-address software sandboxing has the potential to mitigate many more, and to enable a more disruptive shift to stronger compartmentalized software architectures.

Arm has recently developed the *Morello* architecture and processor, incorporating the CHERI protection model into a contemporary high-performance Arm design. Morello is an experimental prototype extending the existing Armv8-A architecture and Neoverse N1 64-bit processor design to support CHERI research and evaluation on the path to eventual productization, and to demonstrate the viability of the CHERI technology using real commercial processes and manufacturing. Extensive software porting is establishing feasibility. Development boards are available for research and prototyping as of early 2022, and are already running significant open-source software stacks, such as an adapted version of the FreeBSD OS and KDE desktop stack running with strong memory safety [2].

In this article we give an overview of CHERI and Morello, and pointers to full discussions elsewhere. It is based largely on material from the "Introduction to CHERI" [3] and "Verified security for the Morello capability-enhanced prototype Arm architecture" [4] technical reports; it does not contain new research results.

Introduction

Memory safety bugs continue to be a major source of security vulnerabilities, responsible, for example, for around 70% of those addressed by Microsoft security updates and around 70% of the high-severity bugs impacting Chromium [5, 6]. Their root causes are well-known legacy design choices and limitations of normal practice that date back to the 1960s and 1970s including: processor architectures that permit one to access any memory location via its plain numeric machine address, protected only via the coarse-grain mechanisms of virtual memory; systems programming languages such as C and C++ that let one do memory accesses without further static or dynamic protection; and test-and-debug development methods that cannot provide high assurance. All of these are baked into the critical systems codebase across the industry, and the result, in today's adversarial environment, is that our codebase inescapably contains many programming errors, which all too often lead to exploitable vulnerabilities.

Many approaches have been developed to improve this situation, from better software engineering processes, through better bug-finding tools and better programming languages, to techniques for machine-checked mathematical proofs of correctness and security. All these are worthwhile, but the legacy investment, the need for systems code to work close to the machine, and the inability of bug-finding to provide high assurance has made it very hard to radically improve conventional systems. Best practice remains an endless Red Queen's Race of identifying and patching security vulnerabilities where one can.

Over the last 10 years, the CHERI project [7] has been exploring a new approach to substantially improve the security of mainstream systems. The basic idea is simple: rather than accessing memory locations via plain numeric machine addresses (and using those to implement C/C++ pointer values), provide architectural support for *unforgeable capabilities* that let one do an efficient permission check at access time (capability systems of various kinds also date back to the 1960s). This, with additional mechanisms, lets one enforce fine-grained memory protection and highly scalable software compartmentalization. Crucially, achieving this memory protection requires only relatively modest changes to architectures and processor designs and relatively minor changes to the sources of existing C/C++ systems software, so there is a real prospect of it becoming widely deployable.

For example, a recent study found that 0.036% lines of code in adapted portions of the Wayland, Qt, and KDE open-source Windows-system and desktop software stack required changes to compile and run with CHERI C/C++ [2].

The academic results for CHERI are encouraging, but achieving such adoption first needs an industry-scale evaluation to demonstrate viability and enable that pull, and that needs a high-performance silicon processor implementation and software stack above it. This is beyond what can be done academically, but hard to justify as a purely commercial project. The 2019–24 UK Research and Innovation (UKRI) Digital Security by Design (DSbD) challenge resolves this chicken-and-egg difficulty with a combined public-sector and industry program (in excess of \$200 million) to build and evaluate such a demonstration platform and support research and development above it [8]. Arm, in a consortium with the University of Cambridge, University of Edinburgh, and Linaro, supported in part by DSbD, has designed and built Morello, a CHERI-enabled prototype architecture, processor, system-on-chip (SoC), development board, and software stacks, extending the Armv8.2-A architecture and the high-performance Neoverse N1 processor [9, 10]. The architecture, emulators, and software toolchains have been available since 2019, and boards began to ship to academic, industrial, and government research labs in early 2022. This will allow evaluation of CHERI mechanisms in a variety of configurations and use cases on a state-of-the-art hardware platform, and paves the way for the potential adoption of CHERI into future production architectures and devices.

CHERI is, most unusually, a *hardware/software/semantics* codesign project. It is centered around a few core principles and architectural ideas that can be instantiated to a range of conventional architectures, explored in detail to date for MIPS, RISC-V, Arm-A, and Arm-M, and sketched for x86. Based on these, there is microarchitectural hardware innovation to implement these efficiently; programming language work so that existing C and C++ code can be rebuilt to gain the additional protection with minimal porting cost; systems software work to demonstrate that OSs, language runtimes, and applications can be adapted to CHERI; and formal semantics and verification work to establish with high confidence that these new architecture variants do indeed provide the intended security properties. The latter includes machine-checked mathematical proofs (in the Isabelle

proof assistant) of key security properties of the complete sequential CHERI-MIPS and Morello ISA specifications [11, 4]—the first time this has been possible for modern production architectures.

CHERI has been designed from the outset to make widespread adoption feasible, but there is a long path and a great deal of work to get any substantial new architectural feature deployed. CHERI has been supported by a sequence of Defense Advanced Research Projects Agency (DARPA) projects, from 2010 onward, by the UK government Engineering and Physical Sciences Research Council (EPSRC) funding, by an EU European Research Council (ERC) Advanced Grant, and by grants and donations from Arm, Google, and others, leading up to the Morello program supported by the UK government DSbD funding, Arm, and others.

CHERI has also been informed by (and builds on) two early research collaborations with the NSA in the early 2000s. First was the collaboration by NSA with Robert Watson at Trusted Information Systems (TIS), centered on Type Enforcement (TE) and OS access control. Robert Watson led the Security-Enhanced Berkley Software Distribution (BSD) and Security-Enhanced Darwin work at TIS, developing and transitioning concepts about operating-system structure, containment, and sandboxing to widespread use through FreeBSD, macOS, iOS, and Junos. At Cambridge, his focus turned to finer-grained software compartmentalization within rather than between applications, leading to the Google-supported development of Capsicum [12], an OS capability framework supporting application decomposition for security, and in turn to the limitations of current processor architectures. A second NSA research collaboration, with Mike Gordon and Anthony Fox at Cambridge, focused on formal modeling of Arm and other ISAs in the HOL4 interactive theorem prover. This work led to the L3 ISA modeling language, which was used in the CHERI-MIPS development and verification, and which was a precursor of our Sail language [13], used for the definitions of RISC-V and CHERI-RISC-V, and in the Morello ISA verification.

An overview of CHERI

The basis of CHERI is a modest set of *architecture extensions* adding hardware representations for capabilities and instructions for manipulating them, integrated into a conventional ISA. Then there are microarchitectural hardware implementation

innovations, developed in our CHERI-MIPS and CHERI-RISC-V FPGA implementations and Arm's Morello silicon implementation; software model and CHERI C/C++ innovations to let software benefit from the new hardware support; CHERI systems software adapting substantial bodies of legacy software; and CHERI formal semantics and verification that increase assurance in the architecture design and inform CHERI C/C++.

CHERI architecture extensions

CHERI extends conventional ISAs, which use machine words to represent language-level integers and pointers, with a new type of hardware-supported data, *unforgeable capabilities* [14, 15, 16, 17]. Capabilities can be used to protect (virtual) addresses intended to be used as code or data pointers—those arising from source-language pointers and also those used in the underlying implementations of language features such as local and global variables, thread-local storage, return addresses, C++ virtual table pointers, and interlibrary linkage. All memory accesses, including loads, stores, and instruction fetch, must be authorized by a capability. As with existing kinds of hardware-supported data (e.g., integers, floats, vectors), capabilities are held in registers and in memory; they are loaded, stored, and manipulated using new *capability-aware instructions*.

In a 64-bit CHERI ISA [18], instead of using simple 64-bit machine-word virtual-address pointer values to access memory, restricted only by the memory management unit (MMU), one can use 128+1-bit capabilities containing a virtual address together with the base and bounds of the memory it can access, and permissions and other metadata. A sophisticated compression scheme lets all this be encoded within the capability with acceptable precision [19]. In turn, having the data within the capability enables a fast access-time check (without any additional lookup), faulting if there is a safety violation. A one-bit tag per register and per each capability-sized and aligned unit of memory, cleared in the hardware by any non-capability write (and not directly addressable), ensures capability integrity by preventing forging. Legacy code compiled without capabilities can still access memory via machine words, but these are restricted by default capabilities held in specific registers, so all accesses can be controlled. This architectural mechanism, along with additional sealed-capability features for secure encapsulation, can be

used by programming language implementations and systems software in many ways.

The ISA design lets code shrink the rights associated with capabilities but never grow them: a *capability monotonicity* property. When any instruction constructs a new capability (except in sealed capability manipulation and exception raising), it cannot exceed the permissions and bounds of the capability from which it was derived. That implies *reachable capability monotonicity*: in any execution of arbitrary code, until execution is yielded to another domain, the set of *reachable capabilities* (those accessible to the current program state via registers, memory, sealing, unsealing, and constructing sub-capabilities) cannot increase.

At boot time, the architecture provides initial capabilities to the firmware, allowing data access and instruction fetch across the full address space. Additionally, all tags are cleared in memory. Further capabilities can then be derived (in accordance with the monotonicity property) as they are passed from firmware to boot loader, from boot loader to hypervisor, from hypervisor to the OS, and from the OS to the application. At each stage in the derivation chain, bounds and permissions may be restricted to further limit access. For example, the OS may assign capabilities for only a limited portion of the address space to the user software, preventing use of other portions of the address space.

Similarly, capabilities carry with them *intentionality*: when a process passes a capability as an argument to a system call, the OS kernel can carefully use only that capability to ensure that it does not access other process memory that was not intended by the user process—even though the kernel may in fact have permission to access the entire address space through other capabilities it holds. This is important as it prevents “confused deputy” problems, in which a more privileged party uses an excess of privilege when acting on behalf of a less-privileged party, performing operations that were not intended to be authorized. For example, this prevents the kernel from overflowing the bounds on a userspace buffer when a pointer to the buffer is passed as a system-call argument.

These architectural properties provide the foundation on which a capability-based OS, compiler, and runtime can implement C/C++-language memory safety and compartmentalization.

Compatibility with current designs has been essential to our approach: ChERI composes well with contemporary architectures, microarchitectures, compiler implementations, OS design, and application structure. The software constructs resting on virtual memory—processes and virtual machines—persist, but are augmented by new mechanisms and structures that support ChERI’s protections.

ChERI Hardware

A principal design goal of the ChERI architecture has been to add new architectural primitives with only limited impact on the overall microarchitecture of contemporary processor and memory-subsystem designs. We have explored potential approaches to integrating ChERI into multiple microarchitectures including our locally developed pipelined Bluespec Extensible RISC Implementation (BERI) 64-bit MIPS core and the Bluespec Piccolo, Flute, and Tooba RISC-V cores; Arm has built on these ideas in Morello. There are two key microarchitectural challenges:

- ▶ **Tagged memory:** Conventional dynamic random-access memory (DRAM) does not support capability tagging. Architecturally, the ChERI protection model does not require a particular implementation of tagging, just that tags be suitably protected and properly coherent with the data they protect. In early designs, we used a simple look-aside tag table stored in DRAM and maintained by the memory controller along with a cache; however, performance analysis revealed a significant DRAM access-rate overhead to this approach. This led us to design a hierarchical tag table able to benefit from the non-uniform distribution of capabilities in memory: inevitably, some pages are rich in capabilities (e.g., stacks, vtable storage); whereas, others are not (e.g., memory mapped files, video and image data) [20]. Another option is to reuse additional metadata storage present in more contemporary double data rate (DDR) designs, including bits available for use in error-correction code (ECC), to hold tags.
- ▶ **Capability compression:** In the absence of capability compression, ChERI capabilities would be four times, rather than two times, the native address size, given the need to store three separate virtual addresses (bottom bound, capability address, and upper bound) as well as metadata.

Bounds compression, which exploits redundancy between these three addresses, is therefore essential to reducing the dynamic memory footprint of pointer-intensive applications (such as language runtimes). Developing a compression scheme that balanced software requirements for precision with microarchitectural efficiency was a significant challenge [19].

Apart from these, CHERI has been designed to avoid changing fundamental design choices in current architectures and microarchitectures—essential elements such as pipeline structure, memory subsystem designs including caches, MMUs, and so on, retain their current structure.

CHERI software model and CHERI C/C++

CHERI capabilities are an architectural primitive that can be used for a variety of software purposes up and down the software stack, with potential uses in firmware and boot loaders, OSs, language runtimes, CHERI-specific compartmentalization libraries, and compiler-generated code for the C and C++ application programs [17, 21, 22, 23, 24, 25]. In our research, we have pursued two central use cases of CHERI capabilities within current C/C++-language software stacks:

- ▶ **Fine-grained memory protection:** By utilizing capabilities instead of integers to implement C/C++ language pointers, and through modest extensions to the OS and language runtime, we have implemented strong and efficient spatial, referential, and (optionally) temporal memory safety for these traditionally memory-unsafe languages.
- ▶ **Scalable software compartmentalization:** Capabilities provide an alternative means to construct the software isolation and controlled communication required to implement compartmentalized software designs. Unlike MMU-based compartmentalization (i.e., implemented using virtual memory), capability-based techniques allow for more granular and scalable data sharing, as well as a single-address-space programming model.

Fine-grained memory protection

The underlying principle in CHERI C and C++ memory protection is to implement pointers (both explicit in the language and implied in the runtime

environment) using capabilities. The main CHERI dialects of C and C++ implement all C/C++ pointer types, as well as all implied pointers (e.g., return addresses, the stack pointer, and so on) using capabilities. This changes the application binary interface (ABI), as pointer size has increased, changing the in-memory layout of data structures, etc., just as 64-bit code has a different ABI from 32-bit code. Because pointers and integers are implemented using different types, additional care must be used so that pointer values retain tags where intended; for example, the C type *uintptr_t* (implemented using the hardware capability type) must be used to hold values that could be integers or pointers, as *long* (implemented using the hardware integer type) has room only for the address portion of a pointer, not its metadata and tag. Overall, however, relatively little code experiences disruption with the introduction of strong memory safety.

To a first approximation, where in classic C/C++ an access by buggy code to arbitrary memory would be undefined behavior with respect to the language standard and, in practice, might lead to an exploitable vulnerability, in CHERI C/C++ such an access will raise a capability exception.

This gives spatial protection; additional work shows how one can also guarantee *temporal safety* (e.g., against reuse-after-free errors) above CHERI architectures, with data so far suggesting reasonable overheads for heap temporal safety.

In addition, we support *hybrid-capability code*, with dialects that implement pointer types using integers by default, interpreted with respect to a global default data capability (DDC) able to address code, globals, heap, and stack(s).

Scalable software compartmentalization

Conventional MMU-based software compartmentalization decomposes larger software applications into components that run in isolated processes, linked only by controlled communication implemented using inter-process communication (IPC). This widely deployed technique, found in applications ranging from Google's Chromium web browser to most Apple iOS applications, limits the impact of software compromise by reducing rights and further attack surfaces available to attackers. This technique is especially important because it provides resilience in the presence of not only exploits for unknown vulnerabilities in known classes (such as buffer overflows), but also

protects against future as-yet undiscovered classes of vulnerability and exploit techniques. However, MMU-based compartmentalization designs impose substantial scalability limits due to utilizing multiple address spaces and page-granularity sharing: the number of compartments and their communication is severely limited, with performance significantly impacted as the number of compartments and their communication grow.

CHERI capabilities permit the construct of isolation and controlled sharing within address spaces, offering potentially greater compartmentalization scalability. Compartments are constructed utilizing closed graphs of capabilities, in which an executing compartment has no access to the resources of other compartments nor broader system memory. Bounds and permissions ensure that capabilities assigned to compartments grant access only to intended resources; monotonicity ensures that these rights cannot be modified to include other resources. Temporal safety is important to ensure that data and capabilities do not improperly leak between compartments when memory is freed and reused.

Switching between compartments can be implemented using one of two architectural mechanisms for controlled non-monotonicity: exception handling, which gives access to additional exception-handling capabilities; or a special jump instruction (CInvoke) that atomically unseals and jumps to a pair of code and data capabilities whose object types match. Both architectural mechanisms provide a means by which available capabilities can be widened, but only when executing previously determined code paths. The jump-based mechanism avoids the cost of exceptions (which is microarchitecturally significant) and avoids the need for a privileged ring transition. It also allows the possibility of more scalable software designs—multiple implementations of domain transition can coexist within a single address space.

The semantics of protection domains and domain transitions are flexible, as these architectural primitives support a variety of potential software uses. These include synchronous function-call-like semantics for domain transition, as well as asynchronous message passing. Compartmentalization models could more resemble libraries in the former case and processes in the latter, depending on the implementation.

CHERI systems software

We have developed a reference software stack for the CHERI architecture exploring several key software design dimensions opened up by capabilities. This work has had a number of aims, including playing an essential part in our hardware-software codesign effort to develop CHERI, to allow evaluation and demonstration of the CHERI approach at scale, and to act as templates for use. The reference stack includes the following components.

- ▶ **CHERI Clang/low-level virtual machine (LLVM)/low-level design (LLD):** An extended version of the Clang/LLVM compiler suite and LLD linker that are able to compile and link hybrid-capability and pure-capability code for multiple CHERI-enabled architectures.
- ▶ **CHERI GDB:** An extended version of the GNU debugger (GDB) that is able to debug hybrid-capability and pure-capability code on multiple CHERI-enabled architectures.
- ▶ **CheriBSD kernel:** An extended version of the FreeBSD OS whose kernel can be compiled either as hybrid-capability or pure-capability code, offering different degrees of kernel memory protection. The CheriBSD kernel is also able to host legacy, hybrid-capability, and pure-capability userspace environments. The pure-capability process environment is known as CheriABI, and is a new OS ABI based on ubiquitous userspace use of architectural capabilities. CheriBSD is also able to offer optional temporal memory safety for (non-stack) allocations in pure-capability userspace applications. CheriBSD's colocated process (or coprocess) model allows multiple userspace processes to coexist safely within a shared virtual address space, using CHERI facilities for fast memory sharing and kernel-free context switching. CheriBSD also implements a CHERI-extended version of FreeBSD's bhyve Type-2 hypervisor, and is able to host CHERI-enabled guest virtual machines.
- ▶ **CheriBSD hybrid userspace:** An extended version of the FreeBSD userspace that is minimally modified to support hybrid-capability code execution, including modest additions to the C runtime (CRT) and system libraries (including libc).

- ▶ **CheriBSD CheriABI userspace:** An extended version of the FreeBSD userspace that supports pure-capability execution with modest further extensions.
- ▶ **CheriBSD applications:** A set of extended applications able to operate in the CheriABI process environment, including integrated FreeBSD programs such as open-source secure shell (OpenSSH), and also third-party applications such as Apple's WebKit, the PostgreSQL database, and X11 window system. These all operate with full spatial, referential, and temporal memory safety.
- ▶ **CheriFreeRTOS:** An extended version of the embedded open-source real-time OS (FreeRTOS) that is able to be compiled as pure-capability code, offering spatial and referential memory safety, as well fine-grained software compartmentalization with fault recovery.
- ▶ **CheriOS microkernel:** An experimental CHERI-specific nanokernel and microkernel illustrating a potential set of design choices available when CHERI is an essential part of an OS design, and its use is maximized. This is a single-address-space, asynchronous message-passing OS intended to support extremely granular compartmentalization side-by-side with strong memory safety.

This software stack demonstrates a number of points in the design space. It explores varying degrees of incorporation and adoption of CHERI protection illustrating CHERI's incremental adoptability properties. It also illustrates architectural neutrality for both the CHERI protection model and software designed for it.

CHERI formal semantics and verification

Formal modeling and verification are essential parts of the CHERI engineering process and of the resulting artifacts. Our CHERI extensions to the 64-bit MIPS and 32/64-bit RISC-V ISAs are defined in the Sail modeling language [13] (with CHERI-MIPS previously in L3 [26]). Sail is a clean, engineer-friendly, first-order imperative language for ISA specification with lightweight dependent types [type-checked using satisfiability modulo theories (SMT)] for static type-checking of bit-vector lengths. It has also been used to give complete sequential ISA definitions for ARMv8-A (automatically translated from the Arm-internal ASL definition [27]) and for RISC-V, and for

ISA semantics integrated with architectural concurrency models. The Sail CHERI architecture models are available as open source.

Sail and L3 are used to generate from these primary models:

- ▶ Reference documentation, automatically incorporated into the CHERI ISA specification [14];
- ▶ Executable ISA-level simulators, in C and OCaml, used as oracles to test hardware against and for software bring-up;
- ▶ Hardware instruction test cases [28], used for hardware testing;
- ▶ SMT-LIB definitions of the architecture, used for SMT checking of intended properties; and
- ▶ Theorem-prover definitions of the architecture, in Coq, Isabelle/HOL, and HOL4, used to state and prove security properties.

Figure 1 illustrates these and how they are used in the CHERI engineering process. Importantly, developing these Sail and L3 definitions did not require expertise in semantics or theorem proving, so the researchers and engineers who would otherwise write a prose/pseudocode architecture document could write and own them. As is familiar from other uses of formal specification, this low-cost activity already brought several benefits, even before any proof work was undertaken.

For CHERI-MIPS and Morello, we have proved that the architecture design does satisfy the intended reachable capability monotonicity property described above—that arbitrary code, if given some initial permissions, cannot increase those during its execution (up to the point of any domain transition) [11, 4]. We have also captured the guarantees one has (and the required assumptions) when executing an untrusted subprogram within a controlled isolation boundary.

These are machine-checked mathematical proofs, using the Isabelle proof assistant. This gives a level of confidence that is not achievable with testing alone—indeed, the highest level of assurance possible.

For formal statements and proofs of security properties, it is always crucial to understand exactly what force they have. We should emphasize that these are properties of the architecture designs, the ISA specifications. It is especially important to ensure that those have the intended properties, as otherwise one would be building vulnerabilities into any

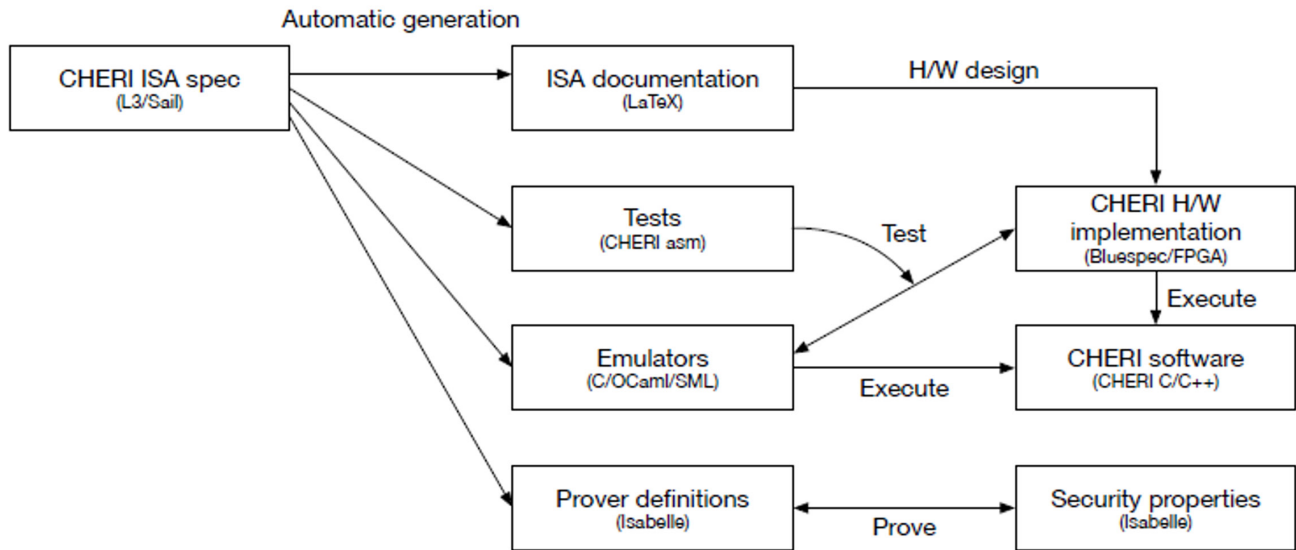


FIGURE 1. The main artifacts of the CHERI-RISC-V and CHERI-MIPS engineering process. Those in the central column are all automatically generated from the Sail (and previously L3) formal ISA specifications. The CHERI hardware design is tested against the generated emulators, using both auto-generated and manually written tests (not shown). The CHERI software stack, including adaptations of Clang and FreeBSD, is developed by running above the generated emulators, the hardware, and a QEMU emulator (not shown). The security properties are stated and proved in terms of the automatically generated Isabelle version of the ISA specification.

conforming hardware implementation (and the proof work did uncover such errors in earlier versions). However, we have not proved the correctness of the CHERI hardware implementations of these architectures (for Morello, that remains beyond the state of the art), or of the software stack above it (though the architecture ensures that considerably less software has to be trusted).

Morello

CHERI has been designed from the outset to be widely deployable, but because it involves changes to the fundamental architectural interface between hardware and software, elaborating and evaluating the approach has required a great deal of work up and down the stack: detailed hardware design, microarchitecture, the architecture design itself, adaptations to the software stack including C/C++ compilers, linkers, debuggers, FreeBSD, and FreeRTOS, and formal semantics, verification, and test generation to ease conventional engineering and improve assurance. It was developed as an academic research project from 2010 onward, but necessarily given the above, a rather large one: with around 100 people involved in one way or another and around \$40 million total funding. All this has produced academically convincing results (reported in many publications) and open-source artifacts.

Achieving widespread adoption of any substantial new architectural feature is challenging, as it needs coordinated hardware and software change across the industry. On the plus side, there are very few architecture vendors, so if a feature becomes (say) part of the mainline Arm architecture and there is pull from major partners, then it will be implemented in all conforming Arm implementations and become ubiquitously available in devices. Around 2018, CHERI faced a chicken-and-egg situation: while the academic results were strong, achieving such adoption needs an industry-scale evaluation to demonstrate viability and enable that pull, and that needs a high-performance silicon processor implementation and software stack above it. But that is beyond what can be done academically and hard to justify as a purely commercial project. The 2019–24 UKRI DSbD challenge, evolved from a 2018 Expression of Interest to the UK Industrial Strategy Challenge Fund, resolves this chicken-and-egg difficulty with a combined public-sector and industry program in excess of \$200 million to build and evaluate such a demonstration platform and support research and development above it [8].

Arm, in a consortium with the University of Cambridge, University of Edinburgh, and Linaro, supported in part by DSbD, has designed and built Morello, a CHERI-enabled prototype architecture,

processor, SoC, development board, and software stacks, extending the Armv8.2-A architecture and the high-performance Neoverse N1 processor [9, 10]. The formal verification of the Morello architecture design, described in the last section, provides assurance that it does provide the intended security properties. The complete CHERI software stack, including compilers, OSs, and applications, have been ported to the architecture. The architecture, emulators, and software toolchains have been available since 2019, and boards have begun to ship to academic, industrial, and government research labs in early 2022. DSbD is also funding further research and evaluation projects centered around Morello. All this will allow evaluation of CHERI mechanisms in a variety of configurations and use cases on a state-of-the-art hardware platform, and paves the way for the potential adoption of CHERI into future production architectures and devices.

CHERI capabilities on Morello

CHERI capabilities are twice the natural address size of the architecture plus an out-of-band tag bit, which is not independently addressable; for Morello, capabilities are 128+1 bits. As shown below, the lower 64 bits are the “value,” which in most cases represents a virtual address. The upper 64 bits encode metadata, including bounds, permissions, and other mechanisms. The tag provides integrity protection: it is preserved only by legitimate operations on capabilities, and cleared by others. A capability can only be used as such (e.g., for a dereference if its tag is set).

perms[17:2]	e	g	otype[14:0]	bounds[86:56]
value[63:0]				

A sophisticated compression scheme allows a capability to include 64-bit lower and upper virtual-address bounds [10, 19]. Small regions can be described precisely, with an arbitrary size in bytes, while for larger regions, only certain bounds and sizes are expressible. The capability value must be either within the bounds or within a certain range above or below, allowing for common C idioms that transiently construct (but do not dereference) slightly out-of-bounds pointers; other combinations of value and bounds are not representable. This scheme trades off bounds precision for reduced capability size; supporting arbitrary bounds would require more than 128+1

bits per capability, which would have unacceptable performance costs.

Four of the 18 permission bits are reserved for software, while the others have architecturally defined meaning. The Load, Store, and Execute permissions control whether a capability can be used for loading or storing data or fetching instructions. Permissions to control loading and storing of capabilities, as opposed to data, are also available. The System permission controls access to system registers and operations, in addition to the access control mechanisms of the base Arm architecture. Capabilities can also be sealed, making them immutable and unusable for anything but branching to them; this allows controlled transitions between different security domains. Sealing (or unsealing) a capability requires an authority capability with the Seal (or Unseal) permission; more on this in the next section.

Capabilities in registers and memory

Morello extends the Armv8-A general-purpose integer register file, as well as certain control and status registers, from 64 bits to 128+1 bits. Memory is extended with a tag bit for each 128-bit sized and aligned unit of DRAM.

The program counter (PC) is extended to become a *program-counter capability* (PCC), constraining instruction fetch as well as PC-relative loads (e.g., of global variables). A new DDC special register controls and transforms memory accesses relative to machine-word pointer values by legacy (non-capability) instructions, for legacy code using integer pointers.

Capability-aware instructions

Morello extends Armv8-A with new instructions and modifies existing instructions to use and respect capabilities. For example, a Load capability (literal) instruction LDR <Ct>,<label> calculates an address from the PCC value and an immediate offset, loads a capability from memory, and writes it to capability register Ct [10]. If the PCC capability does not have the load permission, or the calculated address is outside its bounds, a capability fault exception is raised. The tag of the PCC capability is also checked (already as part of instruction fetching). Most other instructions authorize loads and stores via a capability in an explicitly identified register, or use DDC, rather than implicitly use PCC.

Conventional execution flow is also controlled by capabilities, with branch and branch-and-link-register instructions to capability destinations (or implicitly with respect to the PCC for legacy instructions). Here, too, the capability must have its tag set and the target virtual address must be within the bounds, and in this case, it must authorize execution.

Then there are instructions to access and manipulate the fields of a capability, including arithmetic on its virtual-address value field (corresponding to conventional pointer arithmetic), comparisons, and other operations to extract and manipulate its permissions and other data.

Due to opcode-space constraints, Morello introduces a new instruction-decoding mode that reuses existing integer-relative load/store/jump instructions for capability-relative access.

Domain transition

CHERI distinguishes between sealed and unsealed capabilities. An unsealed capability can be used directly (e.g., to load and store), but a sealed capability can only be used to request actions be taken by other software. This feature can be used in the context of *protection domains* or *software compartments*, in which whole subsystems are given access to a limited subset of memory.

Domain X may have no direct authority to domain Y, but may call into domain Y by *invoking* one or more sealed capabilities originally sealed by (or for) Y. The invocation will install unsealed versions of the invoked capabilities in registers. This always includes replacing the current PCC; thus, this performs a jump to a specific code entry point provided by domain Y. These domain transitions are non-monotonic and must be treated specially in our proof.

Variations on this sealing and invocation mechanism enable slightly different calling styles. When sealing capabilities, they can be labeled with an *object type*, if the authorizing capability has that object type in its bounds. The “branch to sealed capability pair” instruction invokes a given code capability and also an argument data capability, checking their object types match, providing object-style encapsulation. Three kinds of specialized *sentry* (*sealed entry*) capabilities may be used transparently by direct branch instructions, memory-indirect branch instructions, and memory-indirect branch-to-pair instructions, respectively.

Exceptions and the memory management unit

In addition to compiler-facing instructions, system functionality such as virtual memory, cache management, and exception handling is also extended. Exception handling preserves extended capability register state, and there are new exception cause codes associated with CHERI failures such as bounds violations or untagged memory accesses. Because exception handling is able to restore reserved registers during exception-level transitions, it is also a form of domain transition, as reserved registers may contain capabilities not available to the executing code. The page-table format has been extended to add new permissions to, for example, limit loading and storing capabilities.

Conclusion

Work proceeds apace both to explore and evaluate the use of Morello and to investigate the application of the CHERI ideas in other contexts. For further up-to-date details, please see the web pages for CHERI at <https://www.cl.cam.ac.uk/research/security/ctsrds/cheri/> (especially the “Introduction to CHERI” and “Architecture specification” documents there), for Morello at <https://www.arm.com/architecture/cpu/morello>, and for Digital Security by Design at <https://www.dsbd.tech/>. 

Acknowledgements

This work was partially supported by the UK Government Industrial Strategy Challenge Fund (ISCF) under the Digital Security by Design (DSbD) Programme, to deliver a DSbDtech enabled digital platform (grant 105694), ERC AdG 789108 ELVER, EPSRC programme grant EP/K008528/1 REMS, Arm iCASE awards, EPSRC IAA KTF funding, the Isaac Newton Trust, the UK Higher Education Innovation Fund (HEIF), Thales E-Security, Microsoft Research Cambridge, Arm Limited, Google, Google DeepMind, HP Enterprise, and the Gates Cambridge Trust. Approved for public release; distribution is unlimited. This work was supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), under contracts FA8750-10-C-0237 (“CTSRD”), FA8750-11-C-0249 (“MRC2”), HR0011-18-C-0016 (“ECATS”), and FA8650-18-C-7809 (“CIFV”), as part of the DARPA CRASH, MRC, and SSITH research programs. The

views, opinions, and/or findings contained in this report are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

- [1] Joly N, ElSherei S, Amar S. Microsoft Security Response Center (MSRC). "Security analysis of CHERI ISA." 14 Oct 2020. Available at: <https://msrc-blog.microsoft.com/2020/10/14/security-analysis-of-cheri-isa/#:~:text=We've%20assessed%20the%20theoretical,thirds%20of%20all%20those%20issues.>
- [2] Watson RNM, Richardson A, Laurie B. "Assessing the viability of an open-source CHERI desktop software ecosystem." 2021 Sept. Technical Report.
- [3] Watson RNM, Moore SW, Sewell P, Neumann P. "An introduction to CHERI," 2019 Sep. University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-941.
- [4] Bauereiss T, Campbell B, Sewell T, Armstrong A, Esswood L, Stark I, Barnes G, Watson RNM, Sewell P. "Verified security for the Morello capability-enhanced prototype Arm architecture." In: Sergey I. (eds), *Proceedings of the 31st European Symposium on Programming (ESOP) 2022. Lecture Notes in Computer Science, vol 13240*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-99336-8_7.
- [5] Miller M. "Trends, challenges, and strategic shifts in the software vulnerability mitigation landscape." Microsoft Security Response Center (MSRC) BlueHat IL presentation. 2019 Feb. Available at: https://github.com/microsoft/MSRC-Security-Research/blob/master/presentations/2019_02_BlueHatIL/2019_01%20-%20BlueHatIL%20-%20Trends%2C%20challenge%2C%20and%20shifts%20in%20software%20vulnerability%20mitigation.pdf. [Accessed 29 Jun 2021.]
- [6] Chromium Security. "Memory safety." Available at: <https://www.chromium.org/Home/chromium-security/memory-safety>. [Accessed 29 Jun 2021.]
- [7] University of Cambridge, Department of Computer Science and Technology. "Capability Hardware Enhanced RISC Instructions (CHERI)." Available at: <https://www.cl.cam.ac.uk/research/security/ctsr/cheri/>. [Accessed 29 Jun 2021.]
- [8] UK Research and Innovation. "Digital security by design challenge." Available at: <https://www.dsbd.tech/> and <https://www.ukri.org/our-work/our-main-funds/industrial-strategy-challenge-fund/artificial-intelligence-and-data-economy/digital-security-by-design-challenge/>. [Accessed 29 Jun 2021.]
- [9] Arm. "Arm Morello program." Available at: <https://developer.arm.com/architectures/cpu-architecture/a-profile/morello>. [Accessed 29 Jun 2021.]
- [10] Arm. "Arm architecture reference manual supplement Morello for A-profile architecture." Available at: <https://developer.arm.com/documentation/ddi0606/latest>. June 2021. DDI0606A.j. pp. 1288. [Accessed 29 Jun 2021.]
- [11] Nienhuis K, Joannou A, Bauereiss T, Fox A, Roe M, Campbell B, Naylor M, Norton RM, Moore SW, Neumann PG, Stark I, Watson RNM, Sewell P. "Rigorous engineering for hardware security: Formal modelling and proof in the CHERI design and implementation process." In: *Proceedings of the 41st IEEE Symposium on Security and Privacy (SP)*, 2020 May: pp. 1007–1024. doi: 10.1109/SP40000.2020.00055.
- [12] Watson RNM, Anderson J, Laurie B, Kennaway K. "Capsicum: Practical capabilities for UNIX." In: *Proceedings of the 19th USENIX Security Symposium*, 2010: pp. 29–46.
- [13] Armstrong A, Bauereiss T, Campbell B, Reid A, Gray KE, Norton RM, Mundkur P, Wassell M, French J, Pulte C, Flur S, Stark I, Krishnaswami N, Sewell P. "ISA semantics for ARMv8-A, RISC-V, and CHERI-MIPS." *Proceedings of the ACM on Programming Languages*. 2019;3(POPL, 71):1–31. Available at: <https://doi.org/10.1145/3290384>.
- [14] Watson RNM, Neumann PG, Woodruff J, Roe M, Almatary H, Anderson J, Baldwin J, Chisnall D, Davis B, Filardo NW, Joannou A, Laurie B, Marketos AT, Moore SW, Murdoch SJ, Nienhuis K, Norton R, Richardson A, Rugg P, Sewell P, Son S, Xia H. "Capability hardware enhanced RISC instructions: CHERI instruction-set architecture (version 7)." 2019 Jun. University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-927.
- [15] Watson RNM, Neumann PG, Woodruff J, Anderson J, Anderson R, Dave N, Laurie B, Moore SW, Murdoch SJ, Paeps P, Roe M, Saidi H. "CHERI: A research platform deconflating hardware virtualization and protection." In: *Runtime Environments, Systems, Layering and Virtualized Environments (RESOLVE 2012)*, 2012 Mar.
- [16] Woodruff J, Watson RNM, Chisnall D, Moore SW, Anderson J, Davis B, Laurie B, Neumann PG, Norton R, Roe M. "The CHERI capability model: Revisiting RISC in an age of risk." In: *Proceedings of the 41st International Symposium on Computer Architecture (ISCA 2014)*, 2014 Jun: pp. 457–468. doi: 10.1109/ISCA.2014.6853201.

- [17] Watson RNM, Norton RM, Woodruff J, Moore SW, Neumann PG, Anderson J, Chisnall D, Davis B, Laurie B, Roe M, Dave NH, Gudka K, Joannou A, Markettos AT, Maste E, Murdoch SJ, Rothwell C, Son SD, Vadera M. "Fast protection-domain crossing in the CHERI capability-system architecture." *IEEE Micro*. 2016;36(5):38–49. doi: 10.1109/MM.2016.84.
- [18] Watson RNM, Neumann PG, Woodruff J, Roe M, Almatary H, Anderson J, Baldwin J, Barnes G, Chisnall D, Clarke J, Davis B, Eisen L, Filardo NW, Grisenthwaite R, Joannou A, Laurie B, Markettos AT, Moore SW, Murdoch SJ, Nienhuis K, Norton R, Richardson A, Rugg P, Sewell P, Son S, Xia H. "Capability hardware enhanced RISC instructions: CHERI instruction-set architecture (version 8)," 2020 Oct. University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-951.
- [19] Woodruff J, Joannou A, Xia H, Fox A, Norton R, Bauereiss T, Chisnall D, Davis B, Gudka K, Filardo NW, Markettos AT, Roe M, Neumann PG, Watson RNM, Moore SW. "CHERI concentrate: Practical compressed capabilities." *IEEE Transactions on Computers*. 2019;68(10). doi: 10.1109/TC.2019.2914037.
- [20] Joannou A, Woodruff J, Kovacsics R, Moore SW, Bradbury A, Xia H, Watson RNM, Chisnall D, Roe M, Davis B, Napierala E, Baldwin J, Gudka K, Neumann PG, Mazzinghi A, Richardson A, Son S, Markettos AT. "Efficient tagged memory." In: *Proceedings of the 2017 IEEE 35th International Conference on Computer Design (ICCD)*, 2017 Nov: pp. 641–648. doi: 10.1109/ICCD.2017.112.
- [21] Chisnall D, Rothwell C, Davis B, Watson R, Woodruff J, Moore S, Neumann PG, Roe M. "Beyond the PDP-11: Architectural support for a memory-safe C abstract machine." In: *ASPLOS '15: Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2015: pp. 117–130. Available at: <https://doi.org/10.1145/2694344.2694367>.
- [22] Watson RNM, Woodruff J, Neumann PG, Moore SW, Anderson J, Chisnall D, Dave N, Davis B, Gudka K, Laurie B, Murdoch SJ, Norton R, Roe M, Son S, Vadera M. "CHERI: A hybrid capability-system architecture for scalable software compartmentalization." In: *Proceedings of the 36th IEEE Symposium on Security and Privacy*, 2015 May: pp. 20–37. doi: 10.1109/SP.2015.9.
- [23] Memarian K, Matthiesen J, Lingard J, Nienhuis K, Chisnall D, Watson RNM, Sewell P. "Into the depths of C: Elaborating the de facto standards." *ACM SIGPLAN Notices*. 2016;51(6):1–15. Available at: <https://doi.org/10.1145/2980983.2908081>.
- [24] Chisnall D, Davis B, Gudka K, Brazdil D, Joannou A, Woodruff J, Markettos AT, Maste JE, Norton R, Son S, Roe M, Moore SW, Neumann PG, Laurie B, Watson RNM. "CHERI-JNI: Sinking the Java security model into the C." In: *Proceedings of the 22nd ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2017)*; 2017 Apr: pp. 569–583. Available at: <https://doi.org/10.1145/3037697.3037725>.
- [25] Davis B, Watson RNM, Richardson A, Neumann P, Moore S, Baldwin J, Chisnall D, Clarke J, Filardo NW, Gudka K, Joannou A, Laurie B, Markettos AT, Maste E, Mazzinghi A, Napierala ET, Norton R, Roe M, Sewell P, Son S, Woodruff J. "CheriABI: Enforcing valid pointer provenance and minimizing pointer privilege in the POSIX C run-time environment." In: *Proceedings of the 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2019)*; 2019 Apr: pp. 379–393. Available at: <https://doi.org/10.1145/3297858.3304042>.
- [26] Fox ACJ. "Directions in ISA specification." In: Beringer L, Felty A (eds), *Interactive Theorem Proving. ITP 2012. Lecture Notes in Computer Science*, vol 7406. Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/978-3-642-32347-8_23.
- [27] Reid A. "Trustworthy specifications of Arm v8-A and v8-M system level architecture." In: *Proceedings of Formal Methods in Computer-Aided Design (FMCAD 2016)*, 2016 Oct: pp. 161–168. doi: 10.1109/FMCAD.2016.7886675.
- [28] Campbell B, Stark I. "Randomised testing of a microprocessor model using SMT-solver state generation." *Science of Computer Programming*. 2016;118:60–76. Available at: <https://doi.org/10.1016/j.scico.2015.10.012>.



Evaluating Novel Interconnects for Future Cryogenic Computers

Trisha Chakraborty, Laboratory for Physical Sciences (LPS)

Jonathan Cripe, LPS; Department of Physics, University of Maryland, College Park (UMCP)

Karen E. Grutter, LPS

Gregory S. Jenkins, LPS; Department of Physics, UMCP; Quantum Materials Center, UMCP

Kevin D. Osborn, LPS; Department of Physics, UMCP; Quantum Materials Center, UMCP

B. S. Palmer, LPS; Department of Physics, UMCP; Quantum Materials Center, UMCP

Paul Petruzzi, LPS

In this article, we describe the Laboratory for Physical Sciences (LPS) test, evaluation, and research activities associated with the Intelligence Advanced Research Projects Activity's (IARPA) SuperCables program. SuperCables is a program focused on developing the capability to egress single flux quanta (SFQ) data from a superconducting processor at a temperature of 4 Kelvin (K) to room temperature. To achieve the goals of the program, four performers are attempting to map data onto pulses of light and output the data over a thermally insulating optical fiber, instead of the typical way which sends the electrical signal over a conducting microwave cable. To assess the performance of these devices, LPS has developed the capability at 4 K to measure the bit error rates (BERs) up to 30 gigabits per second (Gbps), the dissipated and leaked optical power from nanowatts (nW) up to tens of milliwatts (mW), and has researched the fiber-device coupling, which is a significant source of optical loss in these systems.

[Photo credit: Adobe Stock]



Introduction

The electronic industry has relied on scaling down complementary metal-oxide semiconductor (CMOS) transistors for decades, but the semiconductor-only road map ended in 2015 [1], and the new standard map includes “Beyond CMOS” technologies [2], indicating a need to explore new materials and devices. One alternative to semiconductor logic is superconducting logic [3], which uses single flux quanta (SFQ) for bits [4], and is known for energy efficiency as well as speed. The most developed type, rapid single flux quanta (RSFQ), has been studied as a possible system since the 1990s [5, 6, 7], and modern superconducting logic families show further gains in efficiency by reducing or eliminating static power [8, 9, 10]. Additionally, reversible computing [11, 12, 13, 14] promises the ultimate thermodynamic advancements in digital efficiency, and neuromorphic computing [15, 16, 17] promises new computing architectures, both with superconducting logic. The typical operating temperature for this logic is 4 K (-269 degrees Celsius), and researchers are generally interested in the energy dissipated (i.e., heat produced) during cryogenic operation, which is typically estimated by calculation [18].

Aside from logic circuitry dissipation, the total cryogenic power budget also needs to account for the movement of data (i.e., egress and ingress) between room temperature and 4 K which is significant when transmitting large amounts of data at high data rates

using traditional means. Standard low-loss copper-based microwave cables would transmit a large amount of passive heat from room temperature into the 4 K environment because of the large associated thermal conductivity of the copper (Cu) metal. For example, a standard 50 centimeter-long high-bandwidth microwave cable made from Cu (assuming a temperature thermal conductivity of 3 W/cm*K for a UT-47 cable) would deliver an estimated 0.3 mW of passive power from room temperature directly to the 4 K environment (corresponding to 0.15 W of wall power), a value too large when scaling the number of inputs and outputs for a cryogenic processor. To decrease this amount of heat, one typically uses alloyed metallic cables so that the thermal conductance of heat is reduced, resulting in approximately an order of magnitude decrease in the passive heat. The downsides with this strategy are that resistive cables result in a smaller bandwidth and still pose a challenge for scaling to a very large number of channels of data.

Is there a better way of getting data and signals to and from a cryogenic environment? IARPA's SuperCables program is addressing this question by having performers design and demonstrate devices which convert electrical data from an SFQ stimulus module into pulses of light [i.e., a cryogenic electrical to optical (EO) device], and then use an optical fiber to carry the data signal to room temperature.

There are potentially several benefits to using an optical fiber:

- ▶ Optical fibers enable a much smaller transfer of heat into the cryogenic environment since the thermal conductivity of glass is three orders of magnitude smaller than metal.
- ▶ Data transmitted through an optical fiber has a larger bandwidth than electrical data transmitted in an electrical cable.
- ▶ Optical fibers have better scalability potential when going to a large number of inputs and outputs.

To guarantee that the demonstrated EO devices work better than traditional means for the egress of cryogenic data, IARPA has created metrics centered around small BERs, large bandwidth, and small energy-per-bit devices (see [table 1](#)). To meet these goals, four performer teams are researching cryogenic electro-optic transducers based on tunable ring resonators, a superconducting modulator, or a cryogenic laser [19].

To test and evaluate these unique performer devices, a standard requirement for IARPA's programs, our team has developed the expertise and assembled the facilities needed to evaluate the devices. As we describe in more detail in the following sections, this includes:

1. Measuring BERs of the performer devices up to rates of 10 Gbps using a provided SFQ stimulus module or up to 30 Gbps by driving electrical data from room temperature down to the 4 K environment,
2. Researching and developing new fiber-to-chip couplings for a cryogenic environment, and
3. Researching and developing a cryogenic test platform to measure the dissipated electrical and optical power of the devices at 4 K anywhere in the range of 300 picowatts (pW) to 34 mW.

TABLE 1. IARPA goals for the SuperCables program (The metrics discussed in this article are highlighted in bold.)

Metric	Units	Threshold	Objective	Stretch
Energy per bit at 4 K	aJ/bit	1,000	50	10
Total system energy per bit at room temperature	fJ/bit	65	2	0.2
Bit error rate (BER)	-	10^{-6}	10^{-8}	10^{-10}
Channels per chip area	cm ²	10	100	10,000
Data rate per channel	Gbps	10	50	100
Latency	ns	200	50	10

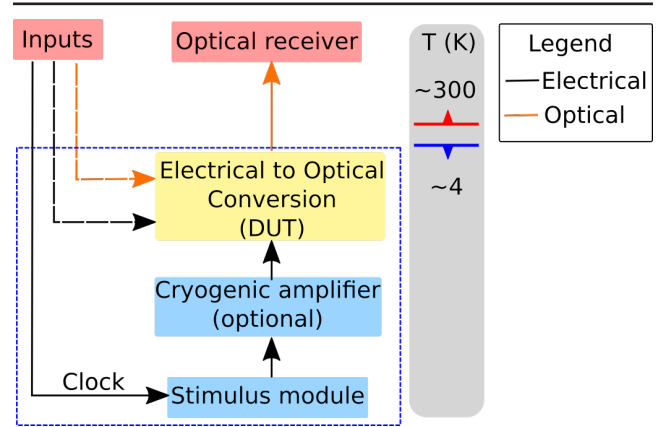


FIGURE 1. In this SuperCables component block diagram, the PRBS pattern is generated at either 300 K using a BERT (dashed black) or at 4 K using a superconducting stimulus module and sent to the electrical input of the EO device under test (DUT) (yellow). The converted optical signal is transported from 4 K to 300 K via an optical fiber (orange) and detected by an optical receiver as part of the BER measurement. Optionally, an input optical signal can be sent in from 300 K to the performer device by a separate optical fiber (dashed orange). The 4 K cryogenic components are shown within the dashed blue box.

Bit error rate measurements

To test the capability of the performer devices to transmit digital data without errors, the SuperCables program specifies the use of bit error rate (BER) measurements. These measurements are performed by a bit error rate tester (BERT) which contains a pattern generator and an error detector. The pattern generator outputs a test pattern, containing a sequence of ones and zeros, that is sent to the input of the performer device. The output of the performer device is then compared to the test pattern by the error detector to determine the number of errors. The BERT result is typically expressed as a ratio, such as 10^{-6} , which means there is one error per million bits. The program goals for BER performance are broken into three tiers including 10^{-6} (threshold), 10^{-8} (objective), and 10^{-10} (stretch).

For the test pattern, the program specifies that we use a pseudorandom bit sequence (PRBS), which is a deterministic, repeating binary pattern that exhibits statistical behavior analogous to a truly random sequence. It is typically denoted by 2^k-1 PRBS or PRBS k , with k being the length of the shift register used to create the pattern. The pattern 2^k-1 PRBS is a sequence that contains every possible combination of the k bits except one. Different length PRBS patterns are used in different applications. For example, 2^7-1 PRBS (127 bits) is often used in Ethernet and Fiber Channel applications [20]. Longer pattern lengths, such as $2^{23}-1$ PRBS (≈ 8 million bits), are typically used for synchronous optical networking (SONET) and synchronous digital hierarchy (SDH) telecommunication systems, which require a pattern with a lower frequency component. For SuperCables, we use a 2^7-1 pattern (see also next section).

A simplified block diagram of the components of our cryogenic BER test bed is shown in figure 1. The electronic PRBS pattern is sent to the performer device at 4 K from either a BERT at 300 K or the SFQ stimulus module, which outputs PRBS and operates adjacent to the device within the cryogenic environment. The converted optical signal is carried to room temperature via an optical fiber and detected by an optical receiver before going to the BERT error detector for analysis. The stimulus module allows us to test the performer device with representative data outputted from a superconducting computer; however, the voltage output amplitude is fixed at a small, approximately 4-millivolt (mV) peak-to-peak level, and the data rates are limited to approximately 10 Gbps, limiting a complete characterization of the devices. On the other hand, our BERT is capable of outputting signals with variable amplitude and at data rates up to 30 Gbps and is therefore able to ascertain a wider range of frequency and amplitude conditions as detailed in the next paragraph.

We conduct a suite of BER measurements for each performer device by sweeping experimental parameters to characterize the device performance and tolerances. Parameters such as optical input power (a laser power) and PRBS amplitude affect the energy efficiency of the device, so the measurements determine how the BER changes as a function of these parameters. As an example, figure 2 shows the measured BER of a commercial optical modulator as a function of optical input power, and while lower optical power ultimately improves the overall energy efficiency, the error rate increases. Error will also be

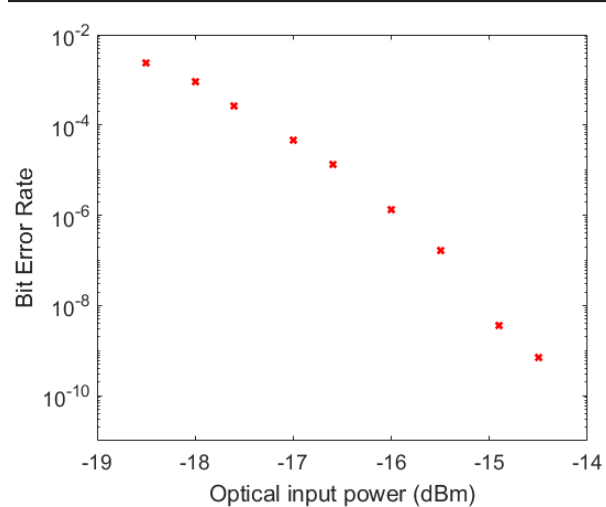


FIGURE 2. The bit error rate (BER) as a function of optical input power is measured here with a commercial optical modulator.

increased for poor optical coupling into the performer devices. Moreover, the photons that do not couple into the device will increase the dissipated heat and reduce the energy efficiency.

When measuring performer devices, we start by taking measurements at room temperature, when possible, and then repeat the measurements at 4 K. We compare the results of our tests to those provided by the performer and determine if they meet the SuperCables objective goals of 10^{-8} . Furthermore, we vary the device bias and measure BER to determine a relationship between dissipated power and BER.

Single flux quantum source

In order to prepare for future tests of performer devices, we cooled and tested the stimulator module. The module uses Niobium (Nb) superconducting circuitry for internal SFQ generation and conversion from SFQ to efficiently made voltage levels which are outputted (as a PRBS waveform). It produces a PRBS signal with a fixed voltage output. This module, shown in figure 3, is mounted on the 4 K plate of the cryostat, as is the performer device, and is connected to the performer device by a short coaxial cable. The module, which can generate either return-to-zero (RZ) or non-return-to-zero (NRZ) patterns, can generate a deterministic PRBS7 or PRBS15 signal with a peak-to-peak amplitude of ~ 4 mV and data rates up to 10 Gbps [21]. While the goal of the program was to have the performer devices driven directly by the stimulus module signal, we also included an optional

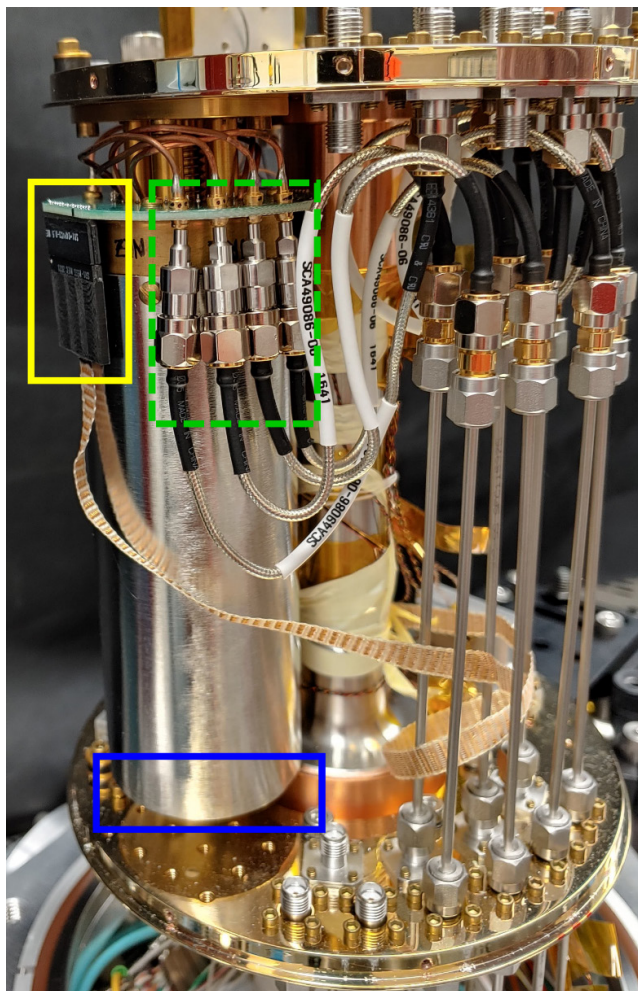


FIGURE 3. In this photo, the stimulus module is mounted to the 4 K plate of the cryostat. The superconducting chip, located at the bottom of the module (not visible, blue box) within the magnetic shield, is attached to the direct current bias and switches (yellow box) and microwave inputs/outputs (dashed green box).

cryogenic amplifier, shown in the schematic of [figure 1](#) to amplify the pattern produced by the stimulus unit as needed. The stimulus module requires 11 direct current biases, four control switches, an input clock signal, and produces two available digital outputs. The module is also encapsulated by two magnetic shields to reduce stray magnetic fields and prevent the trapping of external flux in the niobium (Nb) layers, which would prevent proper operation of the module.

We have successfully operated and characterized the stimulus module between 2.7 K and 4.5 K for data rates up to 10 Gbps. [Figure 4](#) shows a representative PRBS eye diagram generated by the stimulus module, amplified at room temperature and then digitized by an oscilloscope.

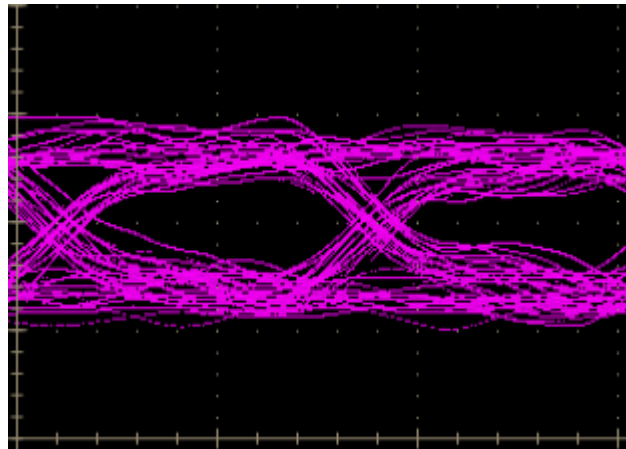


FIGURE 4. This plot shows the output of the SFQ stimulus module operating at 1 Gbps and a temperature of 4 K. The overlaying of repetitive samples produces an eye diagram with the potential to show some of the PRBS signal as ones and zeros. The eye is open in this measurement such that the signal is large compared to noise and the measurement gave a low BER.

Development of low-loss fiber device couplers

One of the biggest challenges in the practical implementation of a cryogenic electro-optic transducer is coupling light between an optical waveguide of the transducer chip and an optical fiber. Typical on-chip optical waveguides are around a few hundred nanometers (nm) wide, whereas the core of a single-mode fiber is around 10 microns (μm) in diameter. Some kind of mode conversion is necessary to efficiently couple between the two. In addition, the method by which the fiber is aligned and affixed to the chip must be cryogenically compatible. Reducing the optical loss due to mode conversion and misalignment is especially important because it not only degrades the performance, but also contributes to heating the cryogenic environment.

Fiber-to-chip coupling strategies generally fall into three categories: edge coupling, grating coupling, and evanescent coupling (see [figure 5](#)). Edge coupling is broadband, polarization-independent, and has shown losses as low as 0.7 decibels (dB) for a single fiber [22], but the alignment tolerance is very low, so thermal contraction from room temperature to the cryogenic environment must be carefully accounted for in the packaging. Grating couplers have better alignment tolerance, but they are typically wavelength-dependent and higher-loss than edge couplers. Evanescent coupling is a less mature

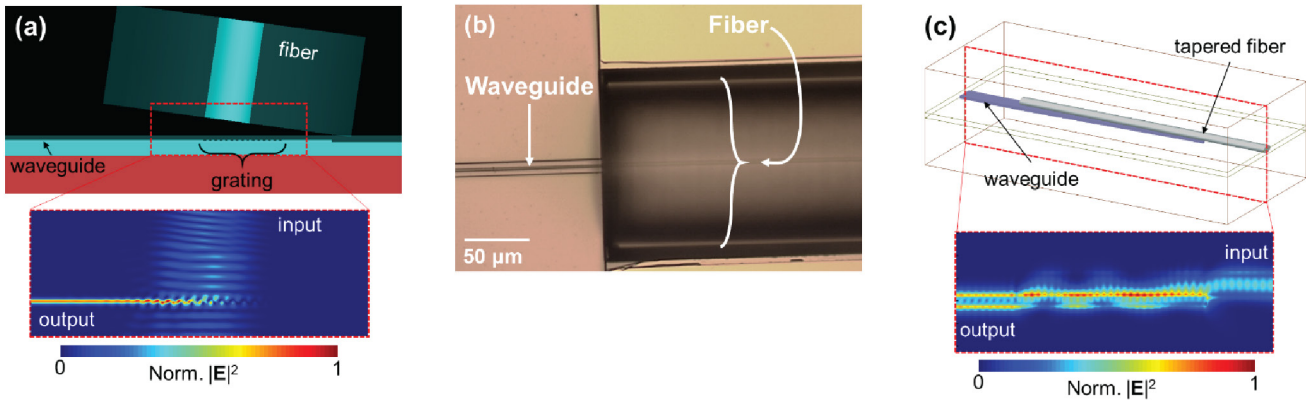


FIGURE 5. The three main strategies for fiber-to-chip coupling are: (a) grating coupling, (b) edge coupling, and (c) evanescent coupling.

coupling strategy but shows similar performance to edge coupling [23]. None of these has emerged as a standard option for packaged cryogenic coupling, partly due to the fact that fibers aligned and affixed at room temperature can move or even break off when cooling down. As a result, most cryogenic electro-optic transducers have been characterized using active alignment with micropositioners in the cryogenic environment, a strategy that is not scalable or power efficient. Here, we describe our work testing epoxy compatibility with the cryogenic environment, then use one of those epoxies in an experiment to test the cryo-robustness of a three-dimensional micron-scale structure for mechanically aligning optical fibers to on-chip devices.

One of the important components for a cryo-robust package is the epoxy used to affix the optical fibers. The physical properties of epoxy will change as it cools, potentially becoming brittle or losing adhesion. As a result, we tested a wide array of epoxies to determine which work best to adhere optical fibers over a wide range of temperatures. We used a sheet of untreated Cu as a substrate and epoxied pieces of optical fiber to the Cu (see figure 6). The epoxies we tested included J-B KwikWeld [24], VGE-7031 [25], Eccobond 286 [26], Silver-impregnated room-temperature-vulcanizing silicon (RTV), rubber cement, and STYCAST 2850 [27]. For each of these, we tried different dilutions with toluene to control the flow around the optical fiber.

We then submerged the sample in liquid nitrogen (77 K) for about 10 minutes to allow it to equilibrate with the nitrogen. After pulling it out, we warmed it with a heat gun. This thermal shock is much greater

than would be experienced in a typical cryostat cooling cycle. We then looked for visible changes in the epoxy, and pulled on the fibers to evaluate adhesion. Most of the epoxies survived without visible changes, but a few did not do well holding the fibers through our pull test. Rubber cement, in particular, did not adhere well to the fiber; it may be useful for other applications, but not for firm fiber attachment.

Having characterized these adhesives, we packaged an integrated photonic chip for cryogenic testing. For this experiment, we leveraged a collaboration with the Lipson Nanophotonics Group at Columbia University to test their “plug-and-play” devices, three-dimensional funnel-like structures that guide

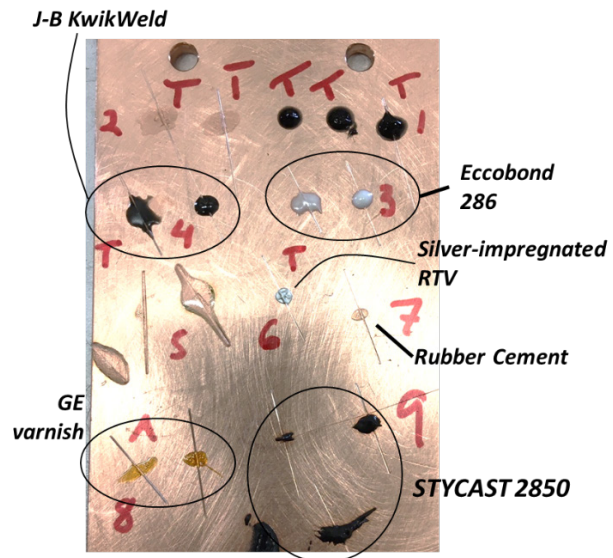


FIGURE 6. The fiber segments are epoxied to copper substrate for a thermal cycling test.

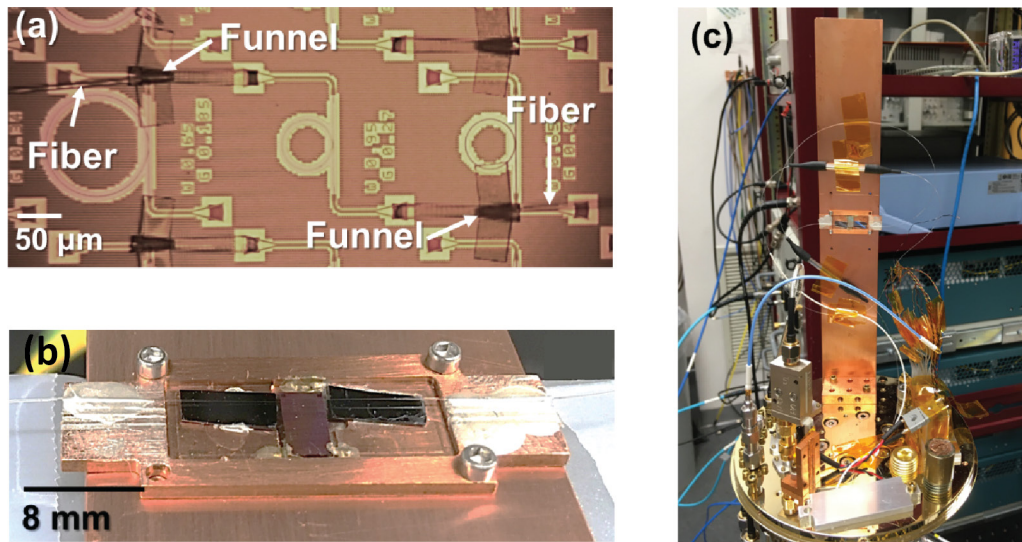


FIGURE 7. (a) In this microscope image of plug-and-play funnels on an integrated photonic chip, fibers are present in the funnels. (b) On this copper mount, chip and fibers are epoxied in place. (c) Here, the chip is mounted inside cryostat.

a tapered optical fiber into alignment with an on-chip grating coupler [see figure 7(a)] [28]. We used silver-impregnated RTV, which survived the thermal shock test particularly well, to hold the fibers in place. A fiber-pigtailed device is shown in figure 7(b).

We mounted the device in a 4 K cryostat provided under the program [figure 7(c)] and monitored the transmission with respect to wavelength as the device cooled. Data from the measurement is shown in figure 8. We were able to see transmission through the device down to the base temperature of the cryostat. Although the maximum transmission did not change very much with temperature, the bandwidth and spectral characteristics shifted slightly, likely due to thermo-optic changes in the refractive index of all three of the materials involved [silicon (Si), silicon dioxide (SiO₂), and the plug-and-play device polymer]. Afterward, we warmed the device back to room temperature.

We repeated this cool-down and warm-up cycle again, and the transmission characteristics did not significantly change. Surviving two cryostat cooling cycles suggests that this packaging strategy is promising as a cryo-robust technique for aligning and affixing fibers to chips. We expect the design can be optimized to further reduce the optical insertion loss. The best reported measurements of these plug-and-play devices was 9.5 dB total insertion loss, with approximately 0.05 dB loss attributable to

each plug-and-play structure (at room temperature) [29]. In addition, the plug-and-play funnel is a versatile, additively-manufactured structure that could be adapted for use with edge couplers and other coupling strategies.

4 K device dissipation measurements

An important metric of the SuperCables program is the 4 K

dissipated energy-per-bit of the performer devices. To directly measure small amounts of dissipated electrical and optical power, our team at LPS has researched and developed new capabilities to measure dissipated powers between 0.3 nW and 30 mW [30].

Figures 9(a) and (b) show pictures and a cross-sectional schematic of our test stage module which allows measurement of heat dissipation. The DUT is bolted to the test stage. Since the module operates in a vacuum, any electrical or optical heat dissipated from the performer device readily conducts into the test stage containing a thermometer and three identical electrical heaters labeled: bias, applied, and feedback heater. The applied electrical current and voltage to each heater is measured to determine the applied heat on the test stage. A thermometer and heater are also mounted to the base plate to actively maintain a stable 3 K cold thermal reservoir.

To simplify discussion of basic heat flow concepts, the test stage with mounted components (heaters, DUT, and thermometer) is hypothetically assumed to be very well thermally isolated. In this idealized scenario, all generated heat that flows into the test stage is energetically stored by raising its temperature. The increase in temperature (ΔT) depends on the total mass (m) of the stage and mounted components. The stored thermal energy is given by $E = (m c_s) \Delta T$, where c_s is the average specific heat of the materials. Measuring small amounts of heat generated from a

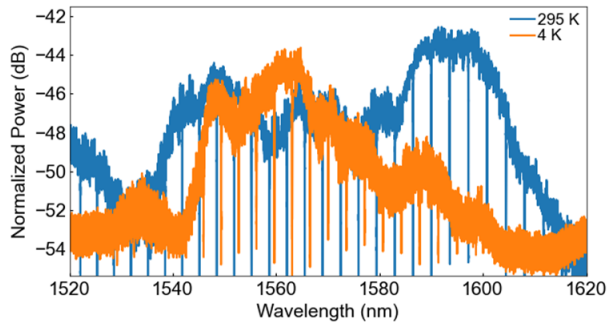


FIGURE 8. Normalized transmission spectrum at room temperature (295 K) and base temperature (4 K).

macroscopic device with large mass is challenging since the signal, which is the rise in temperature, is very small. Rewriting the energy as a heat flow $q = E / \Delta t$ (in watts) and using the definition of heat capacity, $C_s = m c_s$, the instantaneous heat flow into the stage can be written as $q = C_s \partial T / \partial t$. This expression for the storage of heat in material from heat flow is analogous to the storage of charge in a capacitor from a charge flow $I = C \partial V / \partial t$, where current and voltage substitute for heat flow and temperature, and capacitance substitutes for heat capacity.

Connecting the hypothetical test stage to the base plate with a thermally conducting link causes heat to flow from the test stage into the colder base plate. This heat flow depends on the temperature

difference across the thermal link, $\Delta T = q R_L$, where R_L is the thermal resistance of the link that is tuned by selecting its material (thermal conductivity) and adjusting the length and cross-sectional area. A useful electrical analogy is Ohm's law, $\Delta V = I R$, which has electrical resistance R . In order to maintain the test stage at 4 K while tethered to the colder reservoir at 3 K, the test stage bias heater needs to dissipate a power of $q = (4 - 3) K / R_L$ to maintain the temperature difference of 1 K (assuming no power is dissipated by the performer device). Any steady parasitic heat flow into or out of the test stage (from thermal radiation, conduction, and ohmic self-heating of wires and thermometers) is compensated by changing the bias heater power required to maintain the 1 K difference. All parasitic heat leaks between the test stage and base plate are engineered to be much smaller than the heat flow through the thermal link.

A sudden increase in heat flow Δq into the test stage from the performer device that has been turned on will cause a time-dependent rise in temperature away from 4 K. As the temperature rises, added heat energy is stored in the test stage while the temperature difference across the thermal link increases. This results in additional heat flow into the cold reservoir, which reaches a new steady state equilibrium temperature where $\Delta q = \Delta T_{eq} / R_L$. At any time t , the heat flow into the test stage must either go into heating it up or conduct through the link into the cold bath so

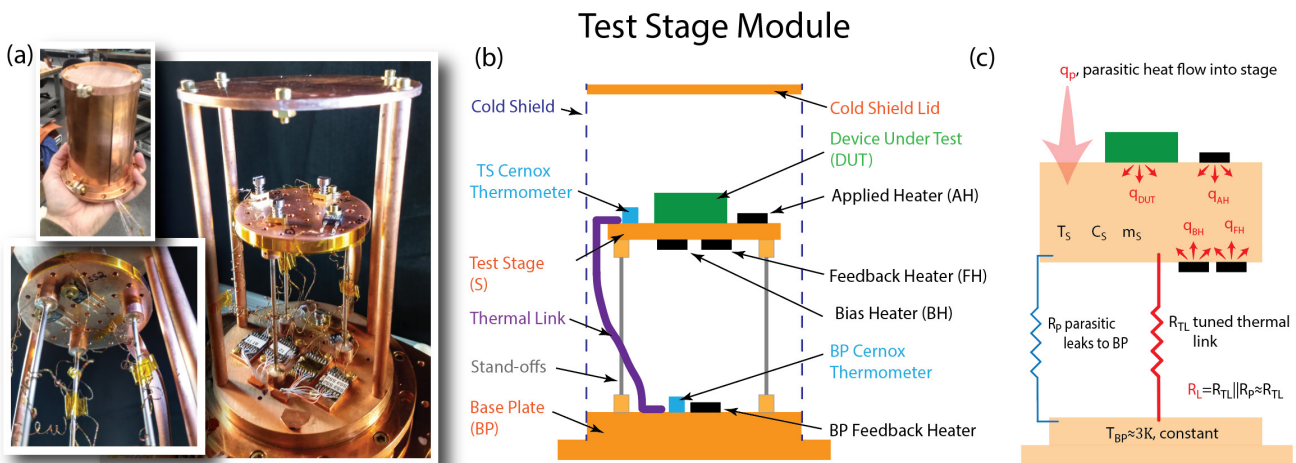


FIGURE 9. (a) These images and (b) this diagram depict the test stage module that measures heat dissipation of cryogenically maintained devices. The module's base plate mounts to a cryostat cold plate in vacuum. A copper test stage is structurally supported by three thin-walled stainless steel tube stand-offs. The test stage is cooled by the base plate through a metal thermal link. (c) A thermal diagram of the test stage module idealizes the test stage as a homogeneous block with a heat capacity C_s , mass m_s and temperature T_s . All test stage heating is from parasitic heating effects q_p plus intentional heating by test devices q_{DUT} and test stage heaters that include the applied heater q_{AH} , feedback heater q_{FH} , and bias heater q_{BH} .

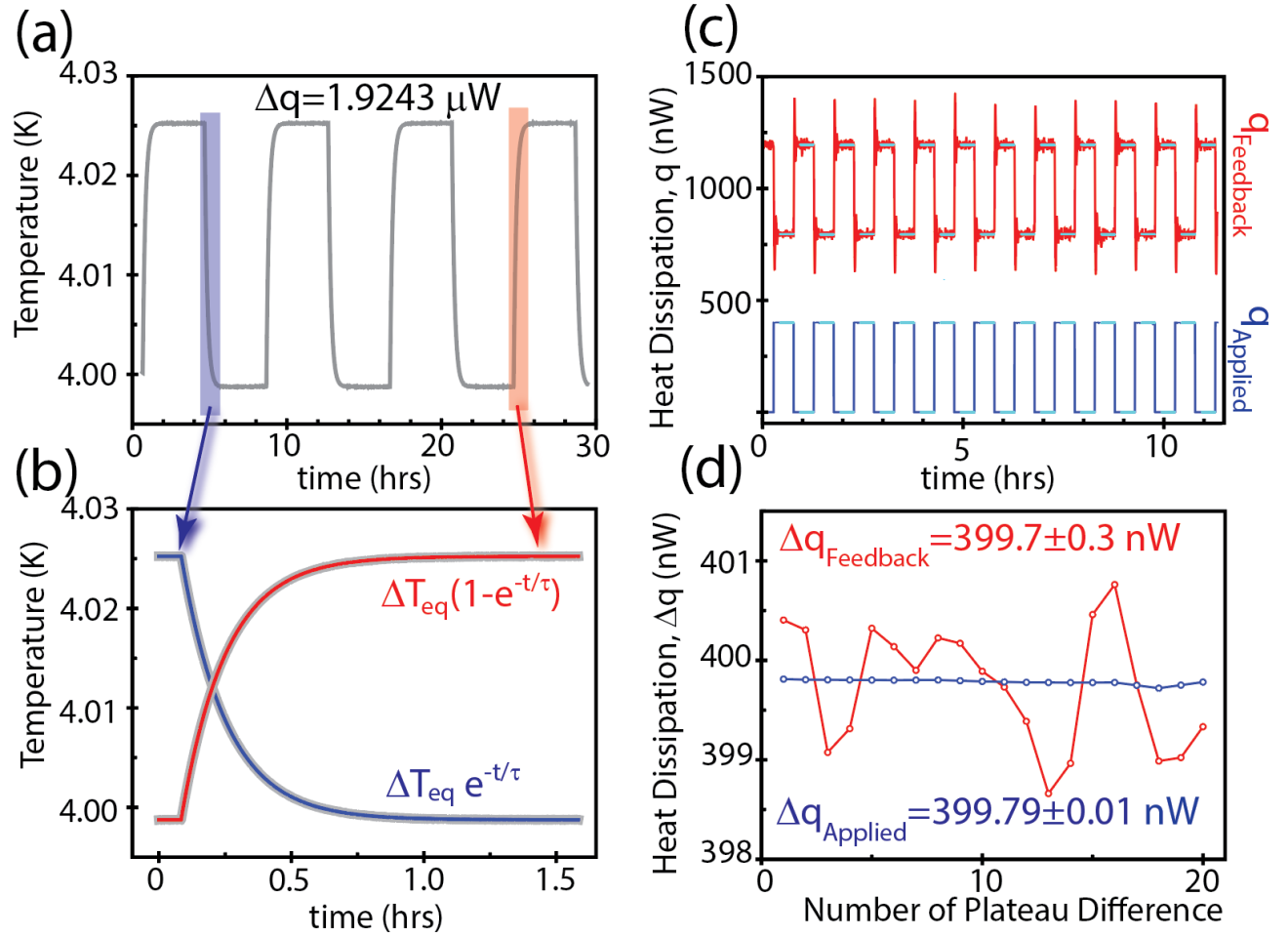


FIGURE 10. (a, b) An applied heater is cyclically toggled off and on (gray) and the temperature response is fit (red and blue) with the parameters ΔT_{eq} and τ , from which R_L and Δq are determined. (c) An active feedback measurement scheme is demonstrated using a weak thermal link. An applied heat of 399.79 nW is cyclically applied (blue) while a feedback heater compensates to maintain the test stage at 4 K (red). The changes in electrical dissipated power Δq are fit (cyan). (d) Each sequential plateau difference Δq_i is plotted for the applied (blue) and feedback heater (red), and the 95% confidence interval of the standard error of the mean reported.

that $\Delta q = C_s \cdot (\partial \Delta T(t) / \partial t) + \Delta T(t) / R_L$. This differential equation has an exponential solution for the time dependence of the test stage temperature and is the same differential equation and solution as a charging capacitor in a series resistor-capacitor (RC) circuit. Multiplying both sides of the equation by resistance, and substituting ΔV for ΔT and ΔI for Δq , gives Kirchhoff's voltage rule. In both cases, the solution has a characteristic RC time constant. For our test stage module, the thermal time constant $\tau = R_L C_s$ is between a few seconds to 10 minutes, depending upon the mass of the device and thermal link selected. The response time is slow when either the resistance is high, which is necessary to produce a measurable rise in temperature when devices dissipate very small heat flows, or when the combined mass of the test stage and mounted components is large.

To characterize the system, we toggle an applied heater off and on and measure the change in temperature of the test stage, as shown in [figure 10\(a\) and \(b\)](#). Exponential fits of the data determine ΔT_{eq} and τ , and therefore R_L . The measured thermal dissipation, found by $\Delta q = \Delta T_{eq} / R_L$, is equal to the measured electrical power supplied to the heater.

Unknown heat loads can be measured by similarly toggling test devices on and off and measuring the temperature rise. However, this measurement method has the inherent shortcoming that devices are measured over a range of temperatures. Thermal conductivity and specific heat of many materials are strongly temperature dependent at cryogenic temperatures, which not only effect R_L and C_s but can also impact device performance.

To circumvent these issues and therefore drastically improve the measurement method, we maintain the test stage and devices at a constant temperature using an active feedback heater. Instead of measuring deviations in the test stage temperature, we measure changes in the electrical power supplied to a feedback heater. While electrical and optical power supplied to the performer device are cyclically toggled on and off, the test stage temperature is measured and maintained at a constant 4 K. The critical engineering considerations using this feedback method remain essentially the same as before. The test stage still responds to changes in heat flow with a characteristic time involving $\tau=R_L C_S$ that critically depends on the mass of the test stage and components. The smallest possible change in heat flow applied by the feedback circuit is predicated upon the smallest measurable temperature deviation as $\delta q_{min}=\delta T_{min}/R_L$, where the maximum thermal resistance R_L (at 4 K) is pragmatically limited by the longest tolerable cool down time of the test stage from room temperature (where the response time τ is much slower).

To demonstrate this feedback method, we use the applied heater to mimic turning off and on the total dissipated power of a performer device with ~ 400 nW as shown in [figure 10\(d\)](#) (blue). The power supplied by the feedback heater compensates (red). Immediately following each toggle event, the response of the feedback heater shows an overshoot followed by rapidly damped oscillations, standard for an optimally tuned proportional integral derivative (PID) feedback circuit. After settling, the feedback heater power plateaus and maintains the test stage at 4 K. Each plateau is averaged (cyan), and averaging the differences between each sequential pair of


plateau values results in a noise floor of 300 pW [see [figure 10\(d\)](#)].

Heat dissipation measurements, using two different thermal links R_L , have been demonstrated using a 430-gram net stage mass of Cu from a minimum of 300 pW up to 30 mW, a dynamic range of 10^8 . Except for the lowest power levels near 300 pW, the precision and accuracy of this technique is measured to four or five significant figures.

Conclusion

Data egress is an important task for the maturation of superconducting computing. With LPS' unique expertise and capabilities in the field of fiber optics, integrated optics, cryogenics, and cryogenic computing, our team has set up a test and evaluation site to evaluate 4 K cryogenic EO devices manufactured under IARPA's SuperCables program. This includes evaluating the error rate of the EO devices by either electrically stimulating it from a superconducting SFQ module or a room temperature data source, measuring the bandwidths of the devices up to 30 Gbps, measuring the dissipation from the devices from 300 pW up to 30 mW, and performing research on the fiber-to-chip coupling.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). The authors would like to thank useful discussions and collaborations with Oscar Jimenez Gordillo. 

References

- [1] HPC Wire. "Transistors won't shrink beyond 2021, says final ITRS report." 2021. *HPC Wire*. Available at: <https://www.hpcwire.com/2016/07/28/transistors-wont-shrink-beyond-2021-says-final-itrs-report/>. [Accessed Sep 2021.]
- [2] IRDS IEEE. International Roadmap for Devices and Systems (IRDS) 2020 Edition. Available at: <https://irds.ieee.org/editions/2020>. [Accessed 8 Sep 2021.]
- [3] Holmes DS, Ripple AL, Manheimer MA. "Energy-efficient superconducting computing—power budgets and requirements." *IEEE Transactions on Applied Superconductivity*. 2013;23(3):1701610–1701610. doi: 10.1109/TASC.2013.2244634.
- [4] Likharev KK, Semenov VK. "RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems." *IEEE Transactions on Applied Superconductivity*. 1991;1(1):3–28. doi: 10.1109/77.80745.

- [5] Likharev KK. "Ultrafast superconductor digital electronics: RSFQ technology roadmap." *Czechoslovak Journal of Physics*. 1996;46:3331–3338. Available at: <https://doi.org/10.1007/BF02548149>.
- [6] Dorojevets M. "Current status and recent developments in RSFQ processor design" In: Luryi S, Xu JM, Zaslavsky A (editors). *Future Trends in Microelectronics: From Nanophotonics to Sensors and Energy*. Wiley & Sons, Inc.; 2010. pp. 229–238. doi: 10.1002/9780470649343.ch19.
- [7] Ishida K, Tanaka M, Nagaoka I, Ono T, Kawakami S, Tanimoto T, Fujimaki A, Inoue K. "32 GHz 6.5 mW gate-level-pipelined 4-bit processor using superconductor single-flux-quantum logic." In: *2020 IEEE Symposium on VLSI Circuits*. 2020 Jun; pp. 1–2. doi: 10.1109/VLSICircuits18222.2020.9162826.
- [8] Kirichenko AJ, Vernik IV, Kamkar MY, Walter J, Miller M, Albu LR, Mukhanov OA. "ERSFQ 8-bit parallel arithmetic logic unit." *IEEE Transactions on Applied Superconductivity*. 2019;29(5):1–7. doi: 10.1109/tasc.2019.2904484.
- [9] Herr QP, Herr AY, Oberg OT, Ioannidis AG. "Ultra-low-power superconductor logic." *Journal of Applied Physics*. 2011;109(10):103903. Available at: <https://doi.org/10.1063/1.3585849>.
- [10] Ayala CL, Tanaka T, Saito R, Nozoe M, Takeuchi N, Yoshikawa N. "MANA: A monolithic adiabatic integration architecture microprocessor using 1.4-z/0p unshunted superconductor Josephson Junction devices." *IEEE Journal of Solid-State Circuits*. 2021;56(4):1152–1165. doi: 10.1109/JSSC.2020.3041338.
- [11] Osborn KD, Wustmann W. "Reversible fluxon logic with optimized CNOT gate components." *IEEE Transactions on Applied Superconductivity*. 2021;31(2):1–13.
- [12] Takeuchi N, Yamanashi Y, Yoshikawa N. "Reversibility and energy dissipation in adiabatic superconductor logic." *Scientific Reports*. 2017;7(1):75. doi: 10.1038/s41598-017-00089-9.
- [13] Ren J, Semenov VK. "Progress with physically and logically reversible superconducting digital circuits." *IEEE Transactions on Applied Superconductivity*. 2011;21(3):780–786. doi: 10.1109/TASC.2011.2104352.
- [14] Frank MP, Lewis RM, Missert NA, Wolak MA, Henry MD. "Asynchronous ballistic reversible fluxon logic." *IEEE Transactions on Applied Superconductivity*. 2019;29(5):1–7. doi: 10.1109/TASC.2019.2904962.
- [15] Shainline JM, Buckley SM, Mirin RP, Nam SW. "Superconducting optoelectronic circuits for neuromorphic computing." *Physical Review Applied*. 2017;7. Available at: <https://doi.org/10.1103/PhysRevApplied.7.034013>.
- [16] Schneider ML, Segall K. "Fan-out and fan-in properties of superconducting neuromorphic circuits." *Journal of Applied Physics*. 2020;128(1). Available at: <https://doi.org/10.1063/5.0025168>.
- [17] Rowlands GE, Nguyen MH, Ribeill GJ, Wagner AP, Govia LCG, Barbosa WAS, Gauthier DJ, Ohki TA. "Reservoir computing with superconducting electronics." 2021. Cornell University Library, arXiv: 2103.02522.
- [18] Takeuchi N, Yamanashi Y, Yoshikawa N. "Measurement of 10 zJ energy dissipation of adiabatic quantum-flux-parametron logic using a superconducting resonator." *Applied Physics Letters*. 2013;102(5). Available at: <https://doi.org/10.1063/1.4790276>.
- [19] Wu H, Fu W, Feng M, Deppe D. "2.6 K VCSEL data link for cryogenic computing." *Applied Physics Letters*. 2021;119(4). Available at: <https://doi.org/10.1063/5.0054128>.
- [20] Maxim Integrated. "Spectral content of NRZ test patterns." Available at: <https://pdfserv.maximintegrated.com/en/an/AN3455.pdf>.
- [21] Lehmann AE, Filippov TV, Sarwana SM, Kirichenko DE, Dotsenko VV, Sahu A, Gupta D. "Embedded RSFQ pseudorandom binary sequence generator for multichannel high-speed digital data link testing and synchronization." *IEEE Transactions on Applied Superconductivity*. 2017;27(4). doi: 10.1109/TASC.2017.2667408.
- [22] Barwicz T, Peng B, Leidy R, Janta-Polczynski A, Houghton T, Khater M, Engelmann S, Fortier P, Boyer N, Green WMJ. "Integrated metamaterial interfaces for self-aligned fiber-to-chip coupling in volume manufacturing." *IEEE Journal of Selected Topics in Quantum Electronics*. 2019;25(3):1–13. doi: 10.1109/JSTQE.2018.2879018.
- [23] Tiecke TJ, Nayak KP, Thompson JD, Peyronel T, de Leon NP, Vuletić V, Lukin MD. "Efficient fiber-optical interface for nanophotonic devices." *Optica*. 2015;2(2):70–75. Available at: <https://doi.org/10.1364/OPTICA.2.000070>.
- [24] JB Weld. KwikWeld. Available at: <https://www.jbweld.com/product/kwikweld-twin-tube>.
- [25] Lake Shore Cryotronics. Varnish. Available at: <https://www.lakeshore.com/products/categories/overview/temperature-products/cryogenic-accessories/varnish>.
- [26] Ellsworth Adhesives. Henkel Loctite Ablestik 286 thermally conductive adhesive white 6 oz kit. Available at: <https://www.ellsworth.com/products/by-market/general-industry/thermally-conductive-materials/adhesives/henkel-loctite-ablestik-286-thermally-conductive-adhesive-white-6-oz-kit/>.

[27] Henkel Adhesive Technologies. Loctite stycast 2850FT. Available at: https://www.henkel-adhesives.com/us/en/product/potting-compounds/loctite_stycast_2850ft.html.

[28] Gordillo OAJ, Chaitanya S, Chang YC, Dave UD, Mohanty A, Lipson M. "Plug-and-play fiber to waveguide connector." *Optics Express*. 2019;27(15):20305–20310. Available at: <https://doi.org/10.1364/OE.27.020305>.

[29] Chakraborty T, Gordillo OAJ, Barrow M, Lipson M, Murphy TE, Grutter KE. "Thermo-optic characterization of SU-8 at cryogenic temperature." In: *Conference on Lasers and Electro-Optics, Technical Digest Series* (Optica Publishing Group, 2022), paper SF30.7. Available at: https://doi.org/10.1364/CLEO_SI.2022.SF30.7.

[30] Jenkins GS, Grutter KE, Petruzzi P, Palmer BS. "Closed cycle 4 K nanowatt meter for hectogram payloads." *AIP Advances*. 2022;12(6):065105. doi: 10.1063/5.0089788.

Next-Generation Radio-Frequency Monitoring in Secure Environments

Minh Nguyen, Brent Laird, Michael R. Gross

Our world is filled with electromagnetic energy, and we are constantly buffeted with both visible and invisible waves radiating throughout our living environment. Electromagnetic energy can be naturally occurring, such as x-rays from black holes and solar flares, shortwave radiation (ultraviolet, visible, and infrared energy) from the sun on a normal day, longwave (mostly infrared) or thermal radiation emitted from the Earth's atmosphere and surface, and radio waves from galactic sources. Electromagnetic energy can also be human-made. The spectrum of energy ranges from extremely low frequency waves such as those emanating from power lines [60 hertz (Hz) in North America [1]] to extremely high frequencies that are used for medical purposes (e.g., diagnostic x-rays, cancer treatments). In between these opposite extremes are visible light and invisible radio frequency [(RF), 3 kilohertz (kHz) to 300 gigahertz (GHz) [2]] energy emitted in the course of many of our day-to-day activities: waking up to a morning radio show, using the microwave to reheat coffee, logging onto a Wi-Fi network, calling a colleague on a cell phone, pairing a fitness tracker using Bluetooth near-field communications, monitoring a home via internet-of-things devices, using a key fob to enter and start a car, using GPS to navigate, using hands-free technology while driving, listening to a satellite radio station, paying a highway toll via radio frequency identification (RFID) transponder, remotely opening a garage door, watching broadcast television, and on and on. In each of these examples, one or more "radios" are transmitting and/or receiving information through electromagnetic waves in the RF. The ubiquity of RF in today's technology results in a constant complex background of RF signals at any given time and place.

[Photo credit: iStock.com/da-kuk]

Should we be worried about all of these RF waves that surround us? Several US government agencies, including the National Institute of Health's National Institute of Environmental Health Sciences [1], Centers for Disease Control [3], Federal Communications Commission (FCC) [2, 4], and the Environmental Protection Agency [5, 6, 7], have reported on research related to potentially adverse health effects of specific portions of the electromagnetic spectrum. To prevent interference across RF signals, particularly in regard to public safety services, the FCC and the National Telecommunications and Information Administration share regulatory responsibility over the allocation of the spectrum between frequency bands 0 kHz and 275 GHz [8]. However, health risks and public safety are not the only concerns posed by this metaphorical ocean of waves. Across the US government, agencies must also consider our complex RF background when identifying potential security risks and implementing methods to mitigate for them. In order to effectively monitor RF signals, they need a system that can sift through enormous volumes of data, in or near real-time, across a wide frequency range. In order to be actionable, such a system must have the ability to differentiate between what is "normal" (in other words, what they would expect to find in the given environment and can thus ignore), what is "anomalous" (unexpected), and what is "significant" (security threat worthy of further investigation).

This article will focus on the last area of concern: security—why we need to monitor signals and what methods and hardware currently exist to meet our needs. Finally, we will introduce new initiatives, including IARPA's Securing Compartmented Information with Smart Radio Systems (SCISRS) program [9], that aim to develop the next-generation methods to automatically detect and characterize suspicious signals and RF anomalies in complex RF environments.

The need to detect anomalous signals

So why are anomalous RF signals a security problem and why would anybody need a system to monitor them? The simplest answer is that certain facilities house highly classified information and, therefore, must have the most rigorous security measures in place. There are strict standards for the physical and technical security of any sensitive compartmented information facility (SCIF) [10, 11]. These standards create a foundation from which those tasked with

securing a SCIF can deduce what normal signals should look like. Attempts to steal or leak data will then give themselves away through telltale signals such as intentional transmissions from unauthorized or modified wireless devices, unexpected mobile cellular signals, and unintentional emanations that carry compromising information. Each type or category of signal has specific methods that are used to detect them. But, as technology progresses, the overall "normal" RF background in secure environments grows more complex, and these indicators of breaches may become easier to hide and more challenging to discover.

US facilities and federal buildings are governed by standards and physical security policies that restrict ingress and egress of items that can receive, record, transmit, or emit information, thereby imposing a technical threat [12]. But the basic equipment necessary to facilitate day-to-day business activities must also be able to receive, record, transmit, or emit information. The varying answers to the simple question "*what is allowed inside of where?*" result in a significant challenge to those responsible for securing these facilities. For example, official electronic devices, information technology (IT), and associated media are permitted if they are operationally required, they have been approved by their organization, and their introduction complies with all relevant policies and procedures. However, the same types of devices (information storage media, radio transmitters, computers, photographic-/audio-/video-recording equipment, and other personal electronic devices) are not allowed, if they do not meet the criteria for official IT. While most personally owned electronic equipment is not permitted in secure facilities, exceptions are made based on a variety of conditions related to the facility (e.g., its location, the types of information housed and exchanged within it) and the capabilities/features of the equipment itself. These might include items needed by individuals with disabilities or for medical reasons (e.g., motorized wheelchairs, hearing aids, pacemakers, electronic hemoglobin-testers, insulin pumps) which are permitted so long as their introductions comply with applicable policies and procedures. Personal cell phones are never permitted inside a SCIF but may be permitted in other parts of the same building. The nature of the policies governing electronic devices and IT result in highly complex RF environments. Now imagine having to do this in less typical surroundings. Some missions require that information and data be generated, stored, used, transmitted, and received in

environments where there is less control. For example, military operations might require a temporary SCIF in order to meet tactical, emergency, or immediate operational requirements. The RF background and baseline would look very different in a remote location versus an urban location, and data security may be more challenging based on variable factors. In some exceptional circumstances, the mandated standards for a SCIF cannot be met, and additional security measures must be taken to mitigate for the increase risk to data security.

Any malicious actors, those whose intent is to steal data, will try to conceal the signals emanating from their devices and by their activities. They may do this by hiding their signals using methods such as spread spectrum or frequency hopping or by employing short bursts that are less likely to be detected in a system that is scanning through a wide spectrum of frequencies; a rough metaphor would be like security guards watching a video feed that jumps from one camera to another and onward until it has completed the circuit of cameras located throughout a building. Vigilant guards would catch illicit activity if it occurred in the time and location that showed up on their monitor, but there is some probability that they could miss it if the activity were very short and fast and there were a large number of cameras to sequentially scan. Alternatively, data thieves may not try to hide their signals at all; rather, they may use signals that mimic or closely resemble those that you would typically see in the target environment. Finally, an opportunist might simply take advantage of unintended RF emanations that inadvertently carry information.

Existing methods to mitigate risks

What can we do to secure our data? Security risks can be binned into broad categories, for which different mitigation strategies are employed. For some categories, the security monitoring is constant; whereas in others, the mitigations are employed on a case-by-case basis. We'll discuss a few examples in the next section.

Wireless intrusion detection system (WIDS)

Rogue wireless devices pose security risks to US facilities, ranging from the nonapproved devices that are inadvertently brought into restricted facilities to hostile devices that intentionally pass information to adversaries. The most minor violations, such as when

an employee forgets a cell phone in a jacket pocket and unintentionally brings it into a SCIF, can result in significant cost to an organization. A typical smartphone contains multiple transceivers including, but not limited to: cellular, Wi-Fi, near-field communications (e.g., Bluetooth), and GPS. Exploitation of cell phones (i.e., interception and monitoring) can enable an adversary to remotely access these transmitters/receivers as well as the phone's storage, camera, and microphone to gain information about the phone's surroundings. Even if a cell phone has not been exploited, accidental introduction into a secure space results in a cost to the organization: forensic analysis of confiscated devices takes time and manpower and can divert critical personnel from mission-critical security duties. Rogue wireless devices can also be *intentionally* introduced by an insider threat—an individual inside an organization who intends to use their authorized access for espionage, unauthorized disclosure of information, or other means of causing damage to the security of the United States. Adversaries can also hide or implant wireless devices inside hardware. Regardless of the intent, the presence of rogue wireless devices impose threats to US information infrastructure.

Over the years, mitigations have been developed and evolved to counter known wireless threats. Because wireless technology is based on communication standards, the detection of unbound RF signals can be used to detect rogue wireless devices. Currently, one of the most common security tools is the wireless intrusion detection system (WIDS), a commercial wireless technology that assists with the monitoring of specific parts of the RF spectrum to identify unauthorized wireless transmissions and/or activities. WIDS can be used to detect, identify, and geolocate wireless local area network (WLAN) devices in controlled spaces. Systems that also include active defense capabilities that can prevent unauthorized connection are wireless intrusion *protection* systems (WIPS). Both WIDS and WIPS consist of an RF sensor component (antennas and radios designed to collect specific wireless transmissions), a central controller/analysis component (software developed to distinguish between authorized/normal and unauthorized/anomalous wireless transmissions), and a display component (the user interface/dashboard that reports findings to designated personnel) [13].

WIDS and WIPS use strategically placed sensors and diagnostic software to track known signals such

as those from Wi-Fi, cellular transmissions, and end-user devices. The components and abilities of commercially available WIDS/WIPS vary based on the manufacturer; however, all systems will have sensors and a server. Hardware-based sensors are comprised of strategically placed antennas paired with radios that are used to scan the relevant channels [typically 2.4 GHz and 5 GHz for Wi-Fi, sometimes 800-900 megahertz (MHz) and 1.8-1.9 GHz for cellular], spending a set amount of time (e.g., 100 milliseconds to 1 second) at each channel. The WIDS/WIPS server detects potential threats by analyzing signatures, behaviors, protocols, and RF spectrum collected by the sensors [14].

Department of Defense (DoD) components deploy WIDS solutions to monitor their controlled spaces for WLAN activity and to detect WLAN-related policy violations on unclassified and classified DoD wired and wireless LANs. WIDS that comply with DoD and other US agency policies [12, 13] are capable of monitoring transmissions that fall within the Institute of Electrical and Electronics Engineers (IEEE) 802.11 body of standards in the 2.4, 3.6, 4.9/5, and 60 GHz spectrum bands. They continuously scan for and detect authorized and unauthorized WLAN activities 24 hours a day, 7 days a week, identifying unauthorized devices interfering with authorized devices, identifying authorized devices operating outside the 802.11 protocol, configuration parameters, and identifying the physical location of all 802.11 devices within the controlled space. In addition to the required 802.11 WLAN protocols, WIDS may also have the capability to detect or monitor traffic of cellular protocols, additional 802.11 protocols, 802.14 protocols, other low-latency protocols, and other long-range wireless protocols.

TEMPEST

Originally a cover name selected by an NSA engineer in the early 1950s, TEMPEST has since become a generic word (noun, verb, or adjective) used in relation to the unintentional emanations of classified information from equipment [15]. Any time a machine is used to process classified information electronically, the various switches, contacts, relays, power lines, and other components may emit electromagnetic or acoustic energy [16]. These emissions behave like small radio broadcasts that radiate through free space, or they may be induced even farther on nearby conductors like signal lines,

external power lines, telephone lines, or water pipes [16]. The potential for an adversary to capture and reconstruct the electromagnetic radiation makes it a security threat. TEMPEST_n, the noun, refers to the technical threat itself; whereas TEMPEST_v, as a verb, can be used to describe the mitigation to reduce the threat, and TEMPEST_{adj}, as an adjective, is used to describe anything related to the phenomenon [15]. The simplest solution to the TEMPEST_{adj} threat is to quantify the distance the TEMPEST_{adj} emanations are able to travel, and establish the zone required to be controlled, and this is the strategy that was adopted by the US military when the problem was first discovered by Bell Telephone during World War II [17]. By 1955, additional techniques were available to suppress TEMPEST_n, and it became possible to TEMPEST_v equipment to prevent it from radiating. In 1976, NSA created the Industrial TEMPEST Program (ITP), a government-industry partnership to develop TEMPEST_n-suppressed equipment to satisfy the government's growing need and reduce the prohibitively high costs for case-by-case mitigations [17]. A few years later, the North Atlantic Treaty Organization (NATO) agreed to a scheme to have vendors offer approved TEMPEST_{adj} products for catalog and sale to NATO and NATO member nations [18]. Despite the successful development of commercially available TEMPEST_n-suppressed equipment, when faced with the need to protect an entire facility housing a large quantity of intelligence-related equipment, an organization might choose to also apply TEMPEST_{adj} countermeasures to the building's construction to shield it in its entirety from TEMPEST_{adj} radiation. Regardless of risk mitigation security measures, TEMPEST_n is a phenomenon that can still be demonstrated and, therefore, a threat that still exists today.

The current state of the art for TEMPEST mitigations can be separated into two categories: prevention and detection. In the first category, the main countermeasures include shielding (putting shields around the equipment to block acoustic or electromagnetic signals), filtering (putting filters on power lines and other outbound connections), masking (structuring devices to emanate signals that don't distinguish between different data values), attenuation (adjusting devices to use less power, minimizing the signal it can radiate), and zoning (establishing a controlled area between equipment and potential adversaries) [19]. In the second category—detection—hardware is used to detect the emanations that could be used to capture and reconstruct

information-bearing signals. Different instrument sensors would be employed in order to capture the different TEMPEST emanations. For example, an oscilloscope may be used to detect voltage signals. A sound transducer could be used to capture acoustic signals. Various antennas and radios would be used to capture other electromagnetic signals in the RF spectrum. Because the wide variety of emanations that can fall under the TEMPEST umbrella, detection methods are often only deployed if there is a specific suspicion of a TEMPEST risk or threat.

Cell-site simulators

A cell-site simulator (also known as fake cell tower, rogue base station, "IMSI catcher," or by commercially available models such as the StingRay) is essentially made up of two components: a software-defined radio (SDR) for sending and receiving radio waves and a computing device to provide a network backend for simulating a cellular core network. Together, they function by transmitting as a cell tower, fooling nearby cellular devices (e.g., cell phones) into identifying the simulator as the best cell tower in the area and subsequently connecting to it. The cell-site simulator receives the unique identifying numbers [international mobile subscriber identity (IMSI)] of those connected devices. When used for criminal justice purposes, law enforcement will use the IMSI to identify its target and obtain signaling information related only to the particular phone that is being targeted [20]. More nefarious actors may connect to any phone, and subsequently perform man-in-the-middle attacks, placing malware between the device and their cellular network, to remove the phone from the real network, clone the target's identity, track location, extract or intercept data, and in some cases deliver spyware [20]. Specialized sensors can be used to detect cell-site simulators. In 2017, the Department of Homeland Security's National Protection and Programs Directorate conducted a limited pilot project that deployed sensors in the National Capitol region in order to identify and better understand potential IMSI catcher activities, and anomalous activity was observed that appeared consistent with IMSI catcher technology including at locations near the White House [22].

Wireless devices, TEMPEST, and cell-site simulators are only a few examples of RF security risks. While current mitigations provide a reasonably high level of confidence in the security of data in

US facilities, these and other threats still exist. Additionally, US data in mobile or temporary environments is more challenging to secure. All of the threats that are described in this article have common elements (signals that are anomalous or unexpected) and similar challenges (their ability to hide in a complex RF background environment).

The next generation of securing information

How do we improve our methods for safeguarding information and data? We know that attempts at data breaches might produce unexpected signals in our known RF environments. The seemingly obvious answer would be to scan all RF signals and analyze them for those that might come from or be used by bad actors. In reality, it would be impractical to install all of the hardware necessary to scan every possible frequency range constantly, and it would be computationally impossible to analyze all of the resulting terabytes of data per second in or near real-time. And our monitoring systems must remain agile to adapt to new evolutions in technology and new signals in our expected RF environment (e.g., 5G millimeter waves). An ideal solution would rely on a balance of efficient algorithms and affordable hardware to reduce the likelihood that an anomalous signal would go undetected.

Let's start with hardware. If the signals of interest are hidden within the expected overt signals and ambient signals that exist normally in the environment, then we need radio receivers to detect all of those signals to analyze. But what kinds of radios? A traditional radio is designed to transmit and/or receive signals in a specific range of frequencies, and the range of frequencies is mainly dependent upon the bandwidth of the radio's antenna and its analog components. For example, that radio in your car most likely receives signals between 540 kHz to 1700 kHz for AM stations and 88 MHz to 108 MHz for FM stations. Remember that the RF spectrum ranges from 3 kHz to 300 GHz (in other words: 3,000 Hz to 300,000,000,000 Hz) which is a very broad range, and suspicious signals may range over several orders of magnitude. If you were limited to traditional analog radios in order to receive signals across the entire RF spectrum, you would need a lot of radios. However, "cognitive radios" or "smart radios" allow us to expand the functionality of radio devices by increasing their frequency spectrum and sampling rates. The FCC has

defined cognitive radio as “a radio that can change its transmitter parameters based on interaction with the environment in which it operates” and cites SDR as an implementation strategy [22]. The FCC goes on to say that “cognitive radio can be viewed as a combined application of SDR and intelligent signal processing with functional elements of radio flexibility, spectral awareness, and the intelligence of decision-making.” SDR uses a small receiver to tune in and listen to radio signals at various frequencies and software to reconfigure itself as needed [24]. SDR is not the radio in and of itself; rather, it is a device that contains a tunable circuit that allows the user, through a software-based tuner, to sample only energy at the desired frequency and sampling rate and ignore all other signals [25]. Much like traditional radios, SDRs rely on antennas, and their utility is limited by the abilities of the antennas they are paired with. Higher-end SDR devices can monitor multiple channels, each providing bandwidth across extended frequency ranges. For example, the Ettus N320 is a networked SDR that has four channels, each providing up to 200 MHz of bandwidth, covering the frequency range from 3 MHz to 6 GHz [26]. Beyond SDRs, we can also use spectrum analyzers to digitize input signals and capture more of a frequency spectrum. The Signal Hound SM200C spectrum analyzer operates in two modes: 1) as a receiver providing in-phase/quadrature phase (I/Q)^a real-time samples with 40 MHz or 160 MHz bandwidth that can be tuned over a 100 kHz to 20 GHz range, or 2) as a spectrum analyzer that sweeps across 100 kHz to 20 GHz at 1 terahertz (THz) per second [27]. Given the right combination of antenna, software, and smart radios, we can receive signals from far more of the RF spectrum with less hardware. And less hardware means less expense.

Assuming we are able to capture a meaningful subset of the RF signals using smart radios, we would need to analyze data at rates approaching terabytes per second. It is a foregone conclusion to say that any analysis at this scale must be automated and most likely will need to employ advanced signal processing (e.g., statistical analysis, analysis of cyclostationary features, machine learning techniques) to be effective. Additionally, because the goal is not only to identify the suspicious signal but to also determine any bad intent and capture the perpetrator, the analysis would need to be completed in near real-time to be actionable. So who is developing this next generation of software algorithms? A number of academic, industry, and government groups are focused on RF

research for varying purposes. The NSA’s Laboratory for Telecommunication Sciences (LTS) is home to an RF research team that investigates, develops, and tests antenna designs and addresses critical challenges, including the detection of RF anomalies [28]. The US DoD’s Defense Advanced Research Projects Agency (DARPA) has invested in RF initiatives in recent years, including, but not limited to: Radio Frequency Machine Learning Systems (RFMLS) to address performance limitations and DARPA Advanced RF Mapping to provide situational awareness which includes the Distributed RF Analysis and Geolocation on Networked System (DRAGONS) project [29, 30]. Other collaborative efforts have been forged between DoD and academia, such as the RF Challenge at the Massachusetts Institute of Technology (MIT), in which the US Air Force has partnered with MIT to fund responses to its challenges, included a Cyber-RF Anomaly Detector Challenge [31]. More specifically to the purposes described in this article, the Intelligence Advanced Research Projects Activity (IARPA) has started a multi-year research effort aimed at developing smart radio techniques that can automatically detect and characterize RF signals potentially associated with attempted data breaches [32].

IARPA, the research and development arm of the Office of the Director of National Intelligence, is the corporate research and development resource for the intelligence community (IC) at large, and it invests in high-risk, high-payoff research programs to tackle some of the most difficult challenges of the agencies and disciplines in the IC. Through its Securing Compartmented Information with Smart Radio Systems (SCISRS) program, IARPA seeks to elevate the IC’s abilities to safeguard information and data that is generated, stored, used, transmitted, and received in secure facilities and beyond [33]. In the fall of 2021, IARPA awarded funding to five performers to develop smart radio techniques to detect and characterize suspicious/anomalous signals in complex RF environments [33]. Over a three-phase 42 month period, the SCISRS performers will develop methods to detect and characterize background and low-probability-of-intercept (LPI) signals such as direct sequence spread spectrum, frequency-hopping spread spectrum, smugglers, and burst (Phase I); altered and mimicked signal anomalies which are signals that resemble known overt signals in frequency, bandwidth, and pulse shape but are unrecognizable to the protocols established to receive them (Phase II); and unintended emissions such as anomalies in

a. In-phase/quadrature phase (IQ) is a mathematical model/representation of a modulated signal.


the emanation baseline arising from microprocessors or other electronics (Phase III) [34].

SCISRS performers will demonstrate the effectiveness of their methods in two test-bed laboratories established and managed by IARPA's collaborative partners. Each test bed houses an operating network, electronic equipment commonly found inside a secure office environment, and other real or synthesized sources required to contribute to both the overt signals and the incidental/unintended RF emissions typically found in an operational environment. While the test beds are located in different geographic areas (Pacific Northwest coast and Mid-Atlantic East coast), their proximity to urban settings, major international airports, radio stations, and other offices provide additional background noise that reflect real-world RF considerations in the two respective geographic locales. In addition to this, anomalous signals (including LPI signals, altered or mimicked signals, and abnormal unintended emissions), with frequencies ranging over several orders of magnitude, will be surreptitiously introduced by the test-bed teams. During the testing periods, performers will be expected to demonstrate their ability to command and control the onsite collection hardware, detect and characterize the ambient signals that make up the RF baseline in the test bed, and perhaps most importantly, characterize and detect anomalous signals that have been added to the ambient baseline.

SCISRS has just begun, and the first phase of testing is anticipated to occur in late 2022/early 2023, with the second and third phases to follow in subsequent 12 month periods. With the completion of each phase, the performers will deliver software to SCISRS repositories. If successful, the initiative will produce the next generation of software algorithms to analyze the massive amounts of data that can be streamed by smart radio systems.

Conclusion

Wired and RF communications systems have faced security threats since the interception of wired telegraph communications during the US Civil War and the later interception of wireless RF communications during the Russo-Japanese War. More recent security risks described in this article continue to persist through the present day. As telecommunications technologies advance, the introduction of more and/or novel signals will present new opportunities

for adversaries. And as the geopolitical landscape changes, new temporary mission-specific secure facilities may be needed. All of these factors, separately or in combination, contribute to the need to grow our abilities to monitor and detect anomalous RF signals. The SCISRS project is poised to deliver novel or improved software solutions to analyze more challenging signals, automate command and control, and potentially provide the means to identify previously undetectable threats. As long as researchers continue to stay vigilant towards future unknown risks, developers target today's known threats, and leaders are open to supporting and adopting new methods, we can continue to secure our nation's most classified information. 

References

- [1] National Institute of Environmental Health Sciences. "Electric and magnetic fields associated with the use of electric power," 2002 Jun. Available at: https://www.niehs.nih.gov/health/materials/electric_and_magnetic_fields_associated_with_the_use_of_electric_power_questions_and_answers_english_508.pdf.
- [2] Federal Communications Commission. "RF safety FAQ frequently asked questions about the safety of radio frequency and microwave emissions from transmitters and facilities regulated by the FCC." Available at: <https://www.fcc.gov/engineering-technology/electromagnetic-compatibility-division/radio-frequency-safety/faq/rf-safety>.
- [3] Centers for Disease Control and Prevention. "Health effects of radiation." 2021. Available at: <https://www.cdc.gov/nceh/radiation/health.html>.
- [4] Federal Communications Commission. "Wireless devices and health concerns." 2020. Available at: <https://www.fcc.gov/consumers/guides/wireless-devices-and-health-concerns>.
- [5] United States Environmental Protection Agency. "Non-ionizing radiation from wireless technology." 2021. Available at: <https://www.epa.gov/radtown/non-ionizing-radiation-wireless-technology#:~:text=Wireless%20technology%20uses%20radiofrequency%20energy,low%2Dlevels%20of%20radiofrequency%20energy>.
- [6] United States Environmental Protection Agency. "Non-ionizing radiation used in microwave ovens." 2021. Available at: <https://www.epa.gov/radtown/non-ionizing-radiation-wireless-technology#:~:text=Wireless%20technology%20uses%20radiofrequency%20energy,low%2Dlevels%20of%20radiofrequency%20energy>.
- [7] United States Environmental Protection Agency. "Electric and magnetic fields from power lines." 2021. Available at: <https://www.epa.gov/radtown/electric-and-magnetic-fields-power-lines>.

- [8] Federal Communications Commission. "Equipment authorization—RF device." Available at: <https://www.fcc.gov/oet/ea/rfdevice#:~:text=The%20FCC%20regulates%20radio%20frequency,9%20kHz%20to%203000%20GHz>.
- [9] Intelligence Advanced Research Projects Agency. SCISRS: Securing Compartmented Information with Smart Radio Systems. Available at: <https://www.iarpa.gov/research-programs/scisrs>.
- [10] Office of the Director of National Intelligence. "Intelligence community standard number 705-1, Physical and technical security standards for sensitive compartmented information facilities." 2010. Available at: <https://www.dni.gov/files/NCSC/documents/Regulations/ICS-705-1.pdf>.
- [11] U.S. General Services Administration. "1025.4 ADM sensitive compartmented information facility use (SCIF) policy." 2020. Available at: <https://www.gsa.gov/directive/sensitive-compartmented-information--facility-use-%28scif%29-policy>.
- [12] Department of Defense. "Department of Defense issuance # DoDI 8420.01 Commercial wireless local-area network (WLAN) devices, systems, and technologies." 2017. Available at: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/842001_dodi_2017.pdf.
- [13] National Security Agency. "Wireless intrusion detection system/wireless intrusion prevention system annex. | version 1.0." 2021. Available at: https://www.nsa.gov/portals/75/documents/resources/everyone/csfc/capability-packages/WIDS-WIPS%20Annex%20v1_0.pdf?ver=u0qF4d82XbjNg8-dNuUuA%3D%3D.
- [14] Coleman DD, Westcott DA, Harkins BE. *CWSP Certified Wireless Security Professional Study Guide: Exam CWSP-205, 2nd Edition ed.* John Wiley & Sons; 2016. ISBN: 978-1-119-21108-2.
- [15] Donahue TM. "Static magic or the wonderful world of TEMPEST or one man's static is another man's treasure!" *Cryptolog*. 1983;10(11). Available at: https://www.nsa.gov/portals/75/documents/news-features/declassified-documents/cryptologs/cryptolog_84.pdf.
- [16] "TEMPEST: A signal problem. The story of the discovery of various compromising radiations from communications and Comsec equipment." *Cryptologic Spectrum Articles*. 1972;2(3). Available at: <https://www.nsa.gov/portals/75/documents/news-features/declassified-documents/cryptologic-spectrum/tempest.pdf>.
- [17] "TEMPEST for every office." *Cryptolog*. 1983;10(11). Available at: https://www.nsa.gov/portals/75/documents/news-features/declassified-documents/cryptologs/cryptolog_84.pdf.
- [18] NATO. "TEMPEST equipment selection process." Available at: <https://www.ia.nato.int/niapc/tempest/certification-scheme>.
- [19] Smith R. *Elementary Information Security, 2nd Edition ed.* Jones & Bartlett Learning Pub; 2015. ISBN-13: 978-1284055931.
- [20] Department of Justice. "Department of Justice policy guidance: Use of cell-site simulator technology," 2015 Sep 3. Available at: <https://www.justice.gov/opa/file/767321/download>.
- [21] Fong M. "Protecting high-level personnel from IMSI catchers." *Security Magazine*. 2020 Feb 21. Available at: <https://www.securitymagazine.com/articles/91767-protecting-high-level-personnel-from-imsi-catchers>.
- [22] Krebs C. Correspondence to Senator Wyden from Christopher Krebs. Available at: <https://www.wyden.senate.gov/imo/media/doc/Krebs%20letter%20to%20Wyden%20after%20May%20meeting.pdf>.
- [23] Federal Communications Commission. "Cognitive radio for public safety," [Online]. Available at: <https://www.fcc.gov/general/cognitive-radio-public-safety>. [Accessed December 2021.]
- [24] Donat W. *Explore Software Defined Radio*. Raleigh (NC): The Pragmatic Bookshelf; 2021. ISBN-13: 978-1680507591.
- [25] Wuff A. *Beginning Radio Communications: Radio Projects and Theory*. Cambridge (MA): Apress; 2019. ISBN-13: 978-1484253014.
- [26] Ettus Research. USRP N320. "Products." Available at: <https://www.ettus.com/all-products/usrp-n320>.
- [27] Signal Hound. "SM200A/B/C spectrum analyzer product manual." 2020. Available at: <https://signalhound.com/sig-downloads/SM200A/SM200-User-Manual.pdf>.
- [28] Laboratory for Telecommunication Sciences. "Research areas." Available at: <https://www.ltsnet.net/research>.
- [29] Davies J. "Radio frequency machine learning systems (RFMLS)." Defense Advanced Research Projects Agency. Available at: www.darpa.mil/program/radio-frequency-machine-learning-systems.
- [30] Rondeau T. "Advanced RF mapping (radio map) (archived)." Defense Advanced Research Projects Agency. Available at: <https://www.darpa.mil/program/advance-rf-mapping>.
- [31] Massachusetts Institute of Technology. "RF Challenge at MIT: Cyber RF anomaly detector challenge." Available at: <https://rfchallenge.mit.edu/challenge-3/>.
- [32] Intelligence Advanced Research Projects Activity. "IARPA announces launch of SCISRS program." 2021 Oct 26. Available at: www.iarpa.gov/newsroom/article/iarpa-announces-launch-of-scisrs-program.
- [33] Office of the Director of National Intelligence. "ODNI news release no. 36-21. IARPA announces launch of SCISRS program." 2021 Oct 26. Available at: <https://www.dni.gov/index.php/newsroom/press-releases/press-releases-2021/item/2257-iarpa-announces-launch-of-scisrs-program>.
- [34] Intelligence Advanced Research Projects Activity. "IARPA-BAA-20-03." 2020 Sep 28. Available at: <https://iarpa.gov/index.php/research-programs/scisrs/scisrs-baa> and <https://sam.gov/opp/f2e9128015684101b2021e04d37516c7/view>.

Detecting Radio-Frequency Electric Fields with Optics

Karen E. Grutter, Laboratory for Physical Sciences (LPS)

Paul Petrucci, LPS

Sumi Radhakrishnan, LPS, Institute for Research in Electronics & Applied Physics, University of Maryland, College Park

Three-dimensional integration (3DI) of integrated circuits (IC) is used today by many chip manufacturers to decrease interconnect length and create greater functionality in a single substrate [1]. These 3DI structures are created by vertically stacking multiple IC die, between which are electrical connections made using through silicon vias (TSV) [2, 3]. This technology brings new failure mechanisms to these chips [4, 5] and new challenges for failure analysis and identification. Typically, scanning electron microscopy (SEM) or x-ray tomography is used to image a damaged chip and determine where the fault occurred. The large size in all three dimensions of a 3DI chip requires techniques to have both a large field of view and high spatial resolution [6]. Since the field of view of these instruments is many times smaller than the IC, it is necessary to stitch together many individual images. This results in extremely long times to acquire the complete image, and stitching errors degrade the image quality. It is possible to eliminate both of these issues by first coarsely localizing the fault and then applying the higher resolution imaging only to this area. Since many of the faults are caused by open or short circuits, it is possible to use electric or magnetic field measurements to provide the coarse fault location. Specifically, causes of open circuits can be a break in an electrical trace or a void in a TSV, to name a few. The voltage drop across the open circuit will create an electric field that could be measured and localized with an electric field sensor with the appropriate spatial resolution and sensitivity.

Detection and measurement of electric fields is traditionally performed using antennas. However, to maximize their sensitivity, the size of a metal antenna is dependent on the wavelength of the field. For radio frequency (RF) electrical fields of 1 megahertz (MHz) to 1 gigahertz (GHz), wavelengths range from 300 meters (m) to 0.3 m, respectively. Although clever antenna design can shrink the size of the antenna to a fraction of these wavelengths, it is still impossible to detect small spatial variations in the field. In addition, the metal used to construct the antenna and the coaxial cable used to connect the antenna to the necessary electronics can interfere with the field that is being measured. In this work, we develop two techniques that are capable of measuring and locating RF electric fields with less than 1 millimeter spatial resolution. The two approaches use optical interferometry which takes advantage of the extreme sensitivity of optical waves to small perturbations and their ability to be confined to small spatial regions, leading directly to the desired spatial resolution.

Overview of optical interferometry

Optical interferometry measures the effective length of an optical resonator using laser light (see [figure 1](#)). If the effective optical length of the resonator changes, either physically (change in length, ΔL) or materially (change in index of refraction, Δn), the laser power transmitted through the cavity will change. Through this mechanism, any change in an outside stimulus can be measured if the relationship between ΔL or Δn of the optical resonator and these stimuli is known. Changes in the effective length of the resonator that are only a fraction of the wavelength of light can be detected, which confers extremely high sensitivity on these sensors. Examples of outside stimuli that have been detected with this phenomenon include temperature [7], pressure [8], acceleration [9], chemical species [10], and gravitational waves [11].

In our case, we can transduce RF fields onto the laser power transmitted through a cavity using special materials that change in response to electric fields. Materials with a linear electro-optic response (or Pockels' effect) exhibit a change in refractive index Δn that is proportional to the electric field, thereby changing the effective length of the optical resonator. Materials with a piezoelectric response change their physical shape in response to an electric field, which can be used to change the physical length L of an optical resonator. The two examples below show each of

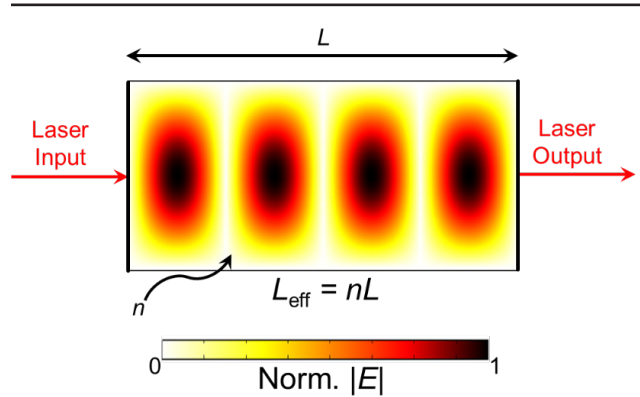


FIGURE 1. Optical interferometry measures the effective length of an optical resonator using laser light; this schematic of an optical resonator has physical length L and index of refraction n .

these material types in action for optical readout of RF electric fields.

The spatial resolution of the electric field measurement is approximately equal to the size of the cavity. Since high-resolution failure analysis techniques, like SEM and x-ray tomography, have a field of view of 100–10 microns (μm), the goal of the two research projects detailed in this paper is to have a spatial resolution in this range. In the two examples below, we will show how both free space and integrated optical sensors achieve spatial resolutions in this range and obtain results that are not possible with antennas.

Electro-optic effect in lithium niobate

Lithium niobate is a human-made material that is not found in nature and, thanks to its extraordinary properties, is commonly used in a wide range of devices including RF filters, acoustic transducers, and optical modulators [12]. In regard to optical devices, lithium niobate is as important as the laser and optical fiber for the incredible capabilities of today's telecommunication networks. Its large electro-optic coefficient combined with a low dielectric constant at optical and RF frequencies allow the use of lithium niobate as an optical modulator in telecommunications networks and, for the same reasons, as an electric field sensor [13]. For an electric field sensor, a laser of wavelength 532 nanometers (nm) is incident on a piece of lithium niobate wafer as shown in [figure 2\(a\)](#). The optical cavity is defined by the laser spot and therefore the spatial resolution is equal to the area of the laser spot. The spot size from a typical laser is on the order of a millimeter but can be reduced to 10 μm or less using a series of lenses.

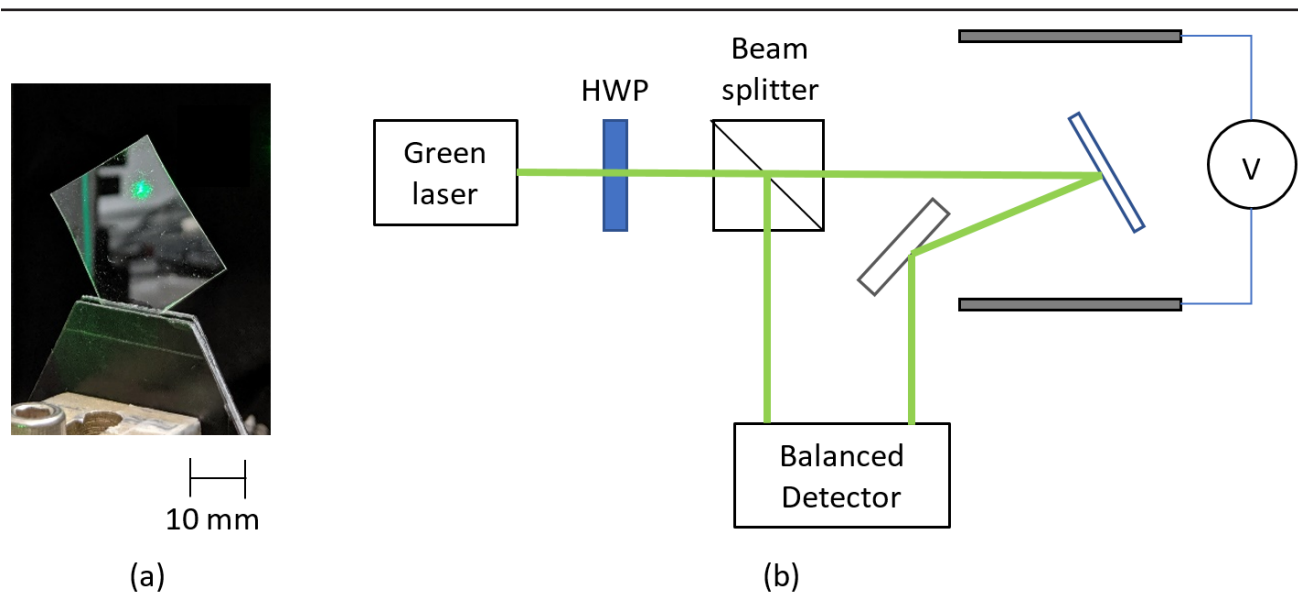


FIGURE 2. (a) In this photograph of a rectangular piece of lithium niobate, a 532 nm laser spot defines the optical cavity of the electric field sensor. (b) This simplified block diagram illustrates the experimental setup for the electric field sensor, HWP = half wave plate.

The setup for characterizing the sensitivity of lithium niobate as an electric field sensor is shown in [figure 2\(b\)](#). Light from a green laser passes through a half-wave plate (HWP) in order to control the angle of the light's polarization relative to the crystal axis of the lithium niobate. There is one specific direction where the electro-optic coefficient is larger than other directions; therefore, to achieve the best electric field sensitivity, the polarization is aligned in this direction. Then the laser passes through a beam splitter that divides the beam into two equal intensity portions. One beam is incident on the lithium niobate sample that is placed between two metal plates, shown in the figure as thin gray rectangles. When a voltage is applied to these plates, a uniform electric field is created with a magnitude equal to the voltage divided by the separation of the plates and a direction perpendicular to the surface of the plates. As with the polarization of the laser beam, the direction of the electric field relative to the crystal axis of the lithium niobate determines the value of the electro-optic coefficient. Therefore, the lithium niobate sample is rotated approximately 45 degrees to align to the direction of the electric field, as shown in [figure 2\(a\)](#). After reflecting off the lithium niobate, the light is directed to a balanced detector where the optical signal is converted to an electrical signal, which is then viewed on an oscilloscope or RF spectrum analyzer. The balanced detector contains two photodiodes that are connected in series, and the output is the center tap between the two. In this way, the currents from

the two photodiodes subtract, thereby canceling the common mode noise between the two optical signals. For this reason, the other output port of the beam splitter is directed to one photodiode of the balanced detector, and the light reflected from the lithium niobate is directed to the other photodiode, thereby significantly reducing the intensity noise from the laser. As will be discussed in the next paragraph, laser intensity noise is a main limiting factor in the sensitivity of this sensor, so reducing it as much as possible is important to improve the performance.

The electric field sensitivity of this sensor is determined by the electro-optic coefficient of the lithium niobate and the noise of the laser. The electro-optic coefficient is 30 picometers per volt (pm/V) which, for a field of 50 V/m, changes the effective length of the interferometer by 8.8×10^{-12} m. This difference in effective length changes the intensity of the light output from the interferometer by 0.01%. Using the balanced detection setup described in the previous paragraph, the measured root mean square (RMS) laser noise is 0.0067%. Therefore, the signal is greater than the noise, and an electric field with amplitude of 50 V/m is measurable using this sensor; the output as recorded by an oscilloscope is shown in [figure 3](#) along with the 1 MHz signal used to create the electric field. Using the measured RMS laser noise, the bandwidth-normalized, minimum detectable electric field is calculated in order to compare the performance of this sensor to other electric field

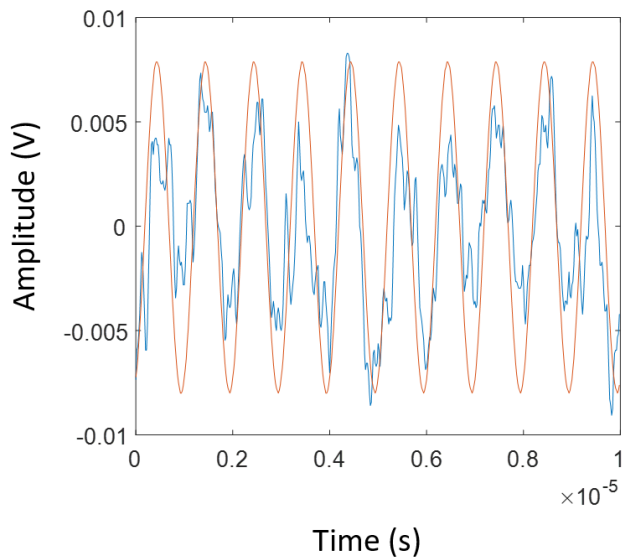


FIGURE 3. In the output signal from the electric field sensor for a 50 V/m electric field at 1 MHz, the blue trace shows the output of the balanced detector captured by an oscilloscope with no averaging, and the red trace is the electric field signal scaled to match the sensor output.

sensors. This minimum detectable electric field is defined as the field required to produce a signal that is equivalent to the noise or a signal-to-noise ratio of one. With this definition, the equation to calculate the sensitivity is

$$\eta = \frac{\sigma_{\text{noise}} \lambda}{\pi n^3 n_{33} L \sqrt{B}}$$

where σ_{noise} is the RMS noise voltage, λ is the laser wavelength, n is the index of refraction, n_{33} is the electro-optic coefficient, L is the thickness of the lithium niobate, and B is the effective noise bandwidth. The values of all the parameters in this equation are shown in [table 1](#) and give a sensitivity of 22 millivolts (mV)/m/Hz^{1/2}. Other lithium niobate electric field sensors have shown sensitivities of 0.13 V/m/Hz^{1/2} [[14](#)] 4.5 V/m/Hz^{1/2} [[15](#)] and 0.35 V/m/Hz^{1/2} [[16](#)].

Piezo-optomechanical nanobeams

One possible method for further increasing sensitivity of these sensors is to measure the interaction between the optical resonance and a mechanical resonance, also known as a cavity optomechanical system. Such systems have been shown to have high sensitivity to displacement; for example, reference [[17](#)] demonstrates a silicon device with displacement sensitivity of approximately 0.5 femtometers

TABLE 1. Parameters and their values used to calculate the sensitivity of the electric field sensor

Parameter	Value
Wavelength (λ)	532 nm
Electro-optic coefficient (n_{33})	30 pm/V
Index of refraction (n)	2.2
Length (L)	1 mm
Effective noise bandwidth (B)	2.5 MHz

(fm)/Hz^{1/2}. In order to couple cavity optomechanical systems to RF signals of interest, we are fabricating them in a piezoelectric material, which deforms in response to an electric field, thus changing the effective optical resonator length.

The device, shown in [figure 4](#), consists of two nanoscale beams with a narrow slot between them [[18](#)]. The top “optical” beam is 42 μm x 745 nm and is a one-dimensional photonic crystal, which has a pattern of elliptical holes designed to confine light in this slot, thereby forming the optical resonance. The bottom “mechanical” beam is 41 μm x 1.1 μm and has a pattern of rectangular holes confining a mechanical “breathing” mode at 2.2 GHz that modulates the width of the narrow slot, thus modulating the optical resonator’s effective length. The breathing mode is not the only mechanical resonance supported by the mechanical beam; [figure 5\(a\)](#) shows several mechanical modes that affect the slot width and thus could be detected optically.

We also simulated the transduction of an ambient RF field into the motion of an aluminum nitride (AlN) beam. For simplicity, we simulated a simple, unpatterned beam, but the results should be qualitatively similar for our mechanical nanobeam [[figure 5\(b\)](#)]. These simulations show that an oscillating electric field can excite the fundamental beam mode, and the amplitude of motion is highest for an electric field in the Z direction (into the page). The absolute displacement will depend on the properties of the deposited material.

To fabricate our AlN double nanobeams, we first used plasma-enhanced chemical vapor deposition to deposit a silicon dioxide (SiO₂) hard mask onto the AlN, which was sputter-deposited onto thermally grown SiO₂. Then, we defined the pattern via electron beam lithography. After etching the pattern into the SiO₂ hard mask with reactive ion etching (RIE), we etched the AlN with inductively-coupled plasma RIE.

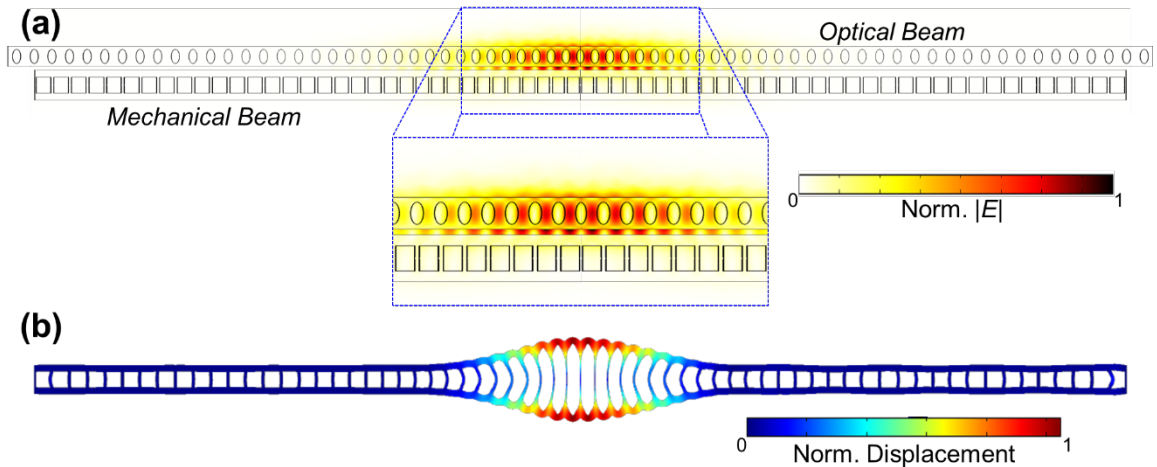


FIGURE 4. (a) This finite element method (FEM) simulation shows the fundamental optical mode at a wavelength of 1,556 nm. **(b)** This FEM simulation shows the mechanical breathing mode at 2.2 GHz. The magnitude of the displacement is exaggerated for clarity.

Finally, we released the beams from the substrate by dissolving the underlying SiO_2 with hydrofluoric acid. An SEM image of a fabricated device is shown in [figure 6\(a\)](#).

To optically characterize our devices, we used a fiber taper waveguide (FTW; minimum diameter of approximately $1 \mu\text{m}$), which allows evanescent coupling of light into and out of the device [[figure 6\(b\)](#)]. We measured several optical resonances at wavelengths from 1,555 nm to 1,599 nm, and they showed intrinsic optical quality factors up to about 45,000 [[figure 6\(c\)](#)]. Using the characterization setup in [figure 7\(a\)](#), we detected mechanical resonances with frequencies ranging from about 10 MHz to 3.6 GHz. Some examples of detected mechanical resonances are shown in [figure 7\(b\)](#) and [7\(c\)](#). Discrepancies in frequency between measured and simulated resonances are likely due to discrepancies between the simulated and

deposited AlN material properties as well as dimensional differences from fabrication variations.

We then brought an RF probe into close proximity with the double nanobeam device while coupling to it optically with a FTW [see [figure 8\(a\)](#)]. The probe consisted of two electrodes separated by $180 \mu\text{m}$, and we centered our device lengthwise between the electrodes, but offset by about $100 \mu\text{m}$, with the probe hovering around 0.5 mm above the surface of the chip. While measuring the optomechanical response around various known mechanical modes, we applied a 0 decibels-per-milliwatt (dBm) RF signal to the RF probe. In addition to the mechanical mode, we were able to see spikes in the spectrum at the frequency of the applied signal [see [figure 8\(b\)](#) and [8\(c\)](#)]. This shows that ambient electric fields can be transduced into mechanical modes of our device, which can then be read out optically. Further work is

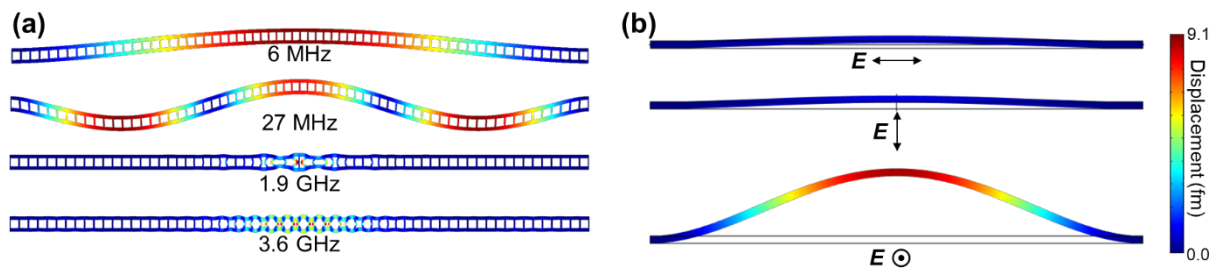


FIGURE 5. (a) These finite element method (FEM) simulations show a selection of resonances supported by the mechanical beam. **(b)** These FEM simulations show a simple aluminum nitride (AlN) beam ($10 \mu\text{m} \times 150 \text{ nm} \times 400 \text{ nm}$) response to a 30 V/m electric field oscillating at 16 MHz in the x , y , and z directions. We assume an isotropic loss factor of 3.2×10^{-4} .

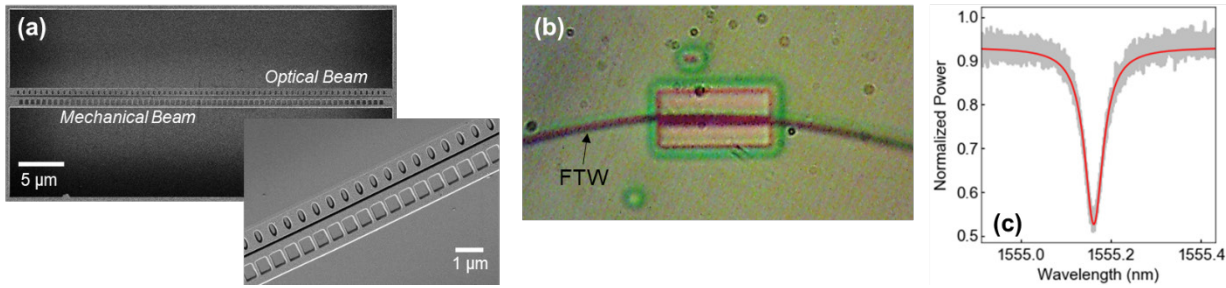


FIGURE 6. (a) This SEM image shows a fabricated AlN double nanobeam. (b) This optical microscope image shows a fiber taper waveguide (FTW) coupling to the device. (c) In this example spectrum of an optical mode, the data is in gray, Lorentzian fit is in red. Intrinsic optical $Q = 42,000 \pm 500$.

required to quantify the polarization dependence and sensitivity of these devices.

Conclusion

In this work, we demonstrated the transduction of ambient RF signals onto optical signals using two different methods. With bulk lithium niobate, we showed the interferometric detection of a 1 MHz signal via the Pockels effect in the material. We also fabricated a nanoscale cavity optomechanical crystal in the piezoelectric material AlN, which we used to detect RF signals at 10 MHz and 1.6 GHz, imprinted on the optical signal at 1,550 nm.

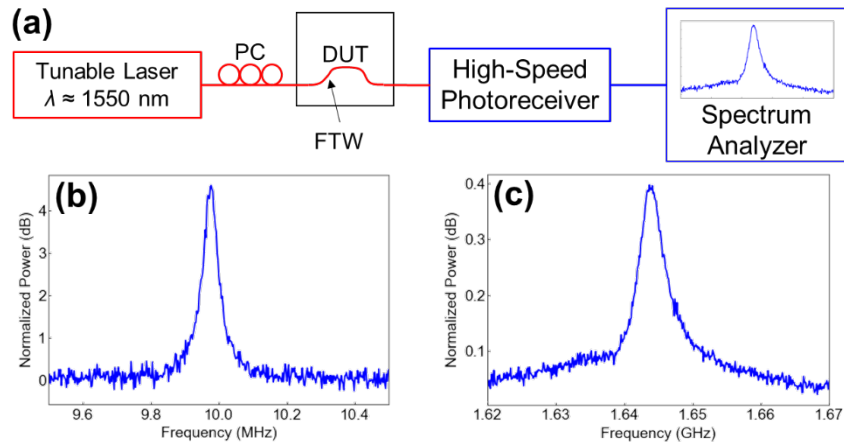


FIGURE 7. (a) Using this characterization setup, we detected (b) a mechanical resonance at 10 MHz (normalized to off-resonance background) and (c) a mode at around 1.6 GHz, normalized to off-resonance background. (PC=polarization controller, DUT=device under test, FTW=fiber taper waveguide.)

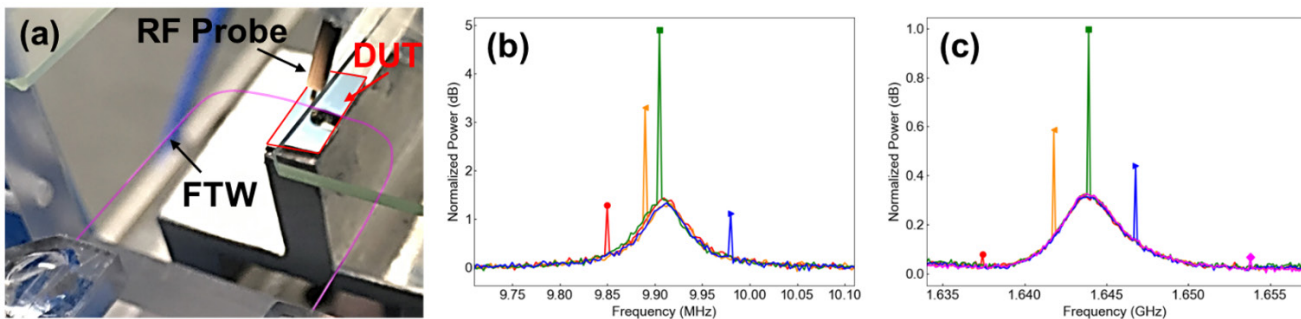



FIGURE 8. (a) In this characterization setup with RF probe hovering over chip, the colored outlines highlight the FTW and the chip containing the device under test (DUT). (b) We detected the spectrum around 10 MHz for different frequency RF excitations, normalized to linear background, and (c) around 1.6 GHz for different frequency excitations. Spectrum is normalized by off-resonance background and slightly offset for best overlay of all spectra. Each colored line corresponds to a different measurement, and the spikes associated with the RF excitations are highlighted with markers.

These types of materials and devices could be used to aid in the detection of faults in IC chips, but the general concept of RF-to-optical transduction via nonlinear material properties can be expanded to a wider range of applications, including microwave-to-optical converters for superconducting computing signal egress and acoustically coupled optical qubits. 

References

- [1] Friedman E, Pavlidis VF. *Three Dimensional Integrated Circuit Design*. Cambridge (MA): Elsevier Inc.; 2017. ISBN: 978-0-12-410501-0.
- [2] Shen WW, Chen KN. "Three-dimensional integrated circuit (3D IC) key technology: Through-silicon via (TSV)." *Nanoscale Res Lett*. 2017;12(56). Available at: <https://doi.org/10.1186/s11671-017-1831-4>.
- [3] Motoyoshi M. "Through-silicon via (TSV)." *Proceedings of the IEEE*. 2009;97(1). doi: 10.1109/JPROC.2008.2007462.
- [4] Choi JW, Guan OL, Yingjun M, Yusoff HBM, Jieli X, Lan CC, Loh WL, Lau BL, Hong LLH, Kian LG, Murthy R, Kiat ETS. "TSV Cu Filling Failure Modes and Mechanisms Causing the Failures." *IEEE Transactions on Components, Packaging and Manufacturing Technology*. 2014;4(4):581–587. doi: 10.1109/TCPMT.2014.2298031.
- [5] Tu KN, Liu Y, Li M. "Effect of Joule heating and current crowding on electromigration in mobile technology." *Applied Physics Reviews*. 2017;4(0011101). Available at: <https://doi.org/10.1063/1.4974168>.
- [6] Orji NG, Badaroglu M, Barnes BM, Beitia C, Bunday BD, Celano U, Kline RJ, Neisser M, Obeng Y, Vladar AE. "Metrology for the next generation of semiconductor devices." *Nature Electronics*. 2018;1(10):532–547. Available at: <https://doi.org/10.1038/s41928-018-0150-9>.
- [7] Purdy TP, Grutter KE, Srinivasan K, Taylor JM. "Quantum correlations from a room-temperature optomechanical cavity." *Science*. 2017; 356(6344):1265–1268. Available at: <https://doi.org/10.1126/science.aag1407>.
- [8] Ma W, Jiang Y, Hu J, Jiang L, Zhang T, Zhang T. "Microelectromechanical system-based, high-finesse, optical fiber Fabry-Perot interferometric pressure sensors." *Sensors and Actuators A: Physical*. 2020;302:111795. Available at: <https://doi.org/10.1016/j.sna.2019.111795>.
- [9] Krause AG, Winger M, Blasius TD, Lin Q, Painter O. "A high-resolution microchip optomechanical accelerometer." *Nature Photonics*. 2012;6:768–772. Available at: <https://doi.org/10.1038/nphoton.2012.245>.
- [10] Ali MM, Memon SF, McGuinness F, Lewis E, Leen G. "Spherical glass based fiber optic Fabry-Perot interferometric probe for refractive index sensing." In: *2020 Conference on Lasers and Electro-Optics (CLEO)*, 2020 May; Jose, CA: pp. 1–2.
- [11] Abbott BP, Abbott R, Abbott TD, Abernathy MR, Acernese F, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al. "Observation of gravitational waves from a binary black hole merger." *Physical Review Letters*. 2016;116(6):061102. doi: 10.1103/PhysRevLett.116.061102.
- [12] Weis RS, Gaylord TK. "Lithium niobate: Summary of physical properties and crystal structure." *Appl. Phys. A*. 1985;37:191–203. Available at: <https://doi.org/10.1007/BF00614817>.
- [13] Cecelja F, Bordovsky M, Balachandran W. "Lithium niobate sensor for measurement of DC electric fields." *IEEE Transactions on Instrumentation and Measurement*. 2001;50(2):465–469. doi: 10.1109/19.918167.
- [14] Rollinson J, Hella M, Toroghi S, Rabiei P, Wilke I. "Thin-film lithium niobate modulators for non-invasive sensing of high-frequency electric fields." *Journal of the Optical Society of America B*. 2021;38(2):336–341. Available at: <https://doi.org/10.1364/JOSAB.412758>.
- [15] Chen L, Reano RM. "Compact electric field sensors based on indirect bonding of lithium niobate to silicon microrings." *Optics Express*. 2012;20(4):4032–4038. Available at: <https://doi.org/10.1364/OE.20.004032>.
- [16] Vohra ST, Bucholtz F, Kersey AD. "Fiber-optic dc and low-frequency electric-field sensor." *Optics Letters*. 1991;16(18):1445–1447. Available at: <https://doi.org/10.1364/OL.16.001445>.
- [17] Srinivasan K, Miao H, Rakher MT, Davanco M, Aksyuk V. "Optomechanical transduction of an integrated silicon cantilever probe using a microdisk resonator." *Nano Letters*. 2011;11(2):791–797. Available at: <https://doi.org/10.1021/nl104018r>.
- [18] Davanco M, Chan J, Safavi-Naeini AH, Painter O, Srinivasan K. "Slot-mode-coupled optomechanical crystals." *Optics Express*. 2012;20(22):24394–24410. Available at: <https://doi.org/10.1364/OE.20.024394>.

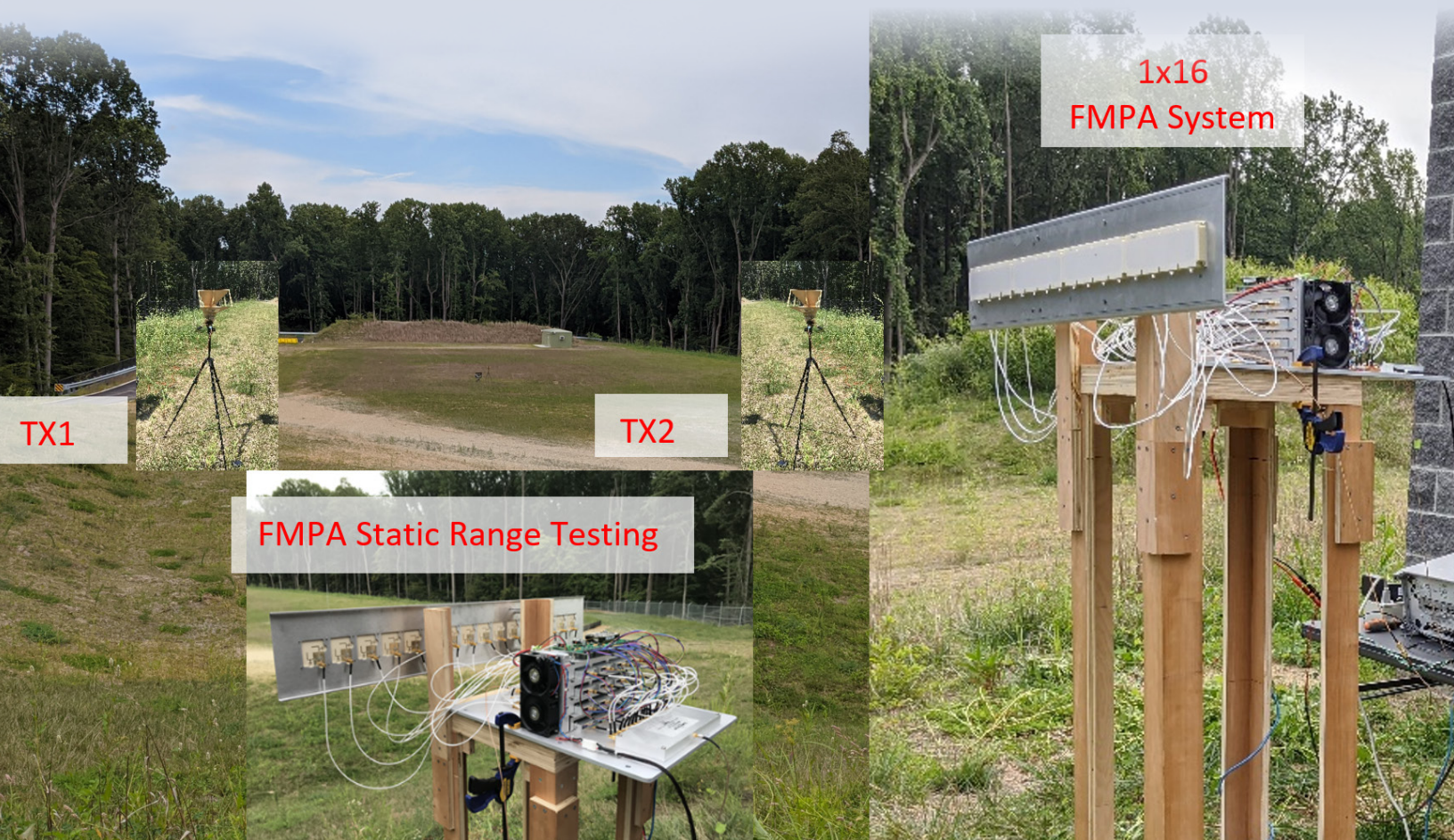


00 110 120 130 140 150

A Novel Hardware Concept for Digital Beamforming: Development and Testing of a Frequency Multiplexed Phased Array (FMPA) System

David Elsaesser
Laboratory for Physical Sciences

Spyro Gumas, Ravi Goonasekeram, Timothy Sleasman, John Marks
Johns Hopkins University Applied Physics Laboratory



In this article, we propose a novel frequency multiplexing approach to digital beamforming using an identical frequency downconverter (DC) for each array antenna element and a commercial-off-the-shelf single-channel digital receiver. We compare this technique to coherent multi-channel/multi-receiver-based architectures to establish the benefits of the multiplexing approach. We build and test an S-band prototype system to prove that the frequency multiplexing concept is viable. Development of this prototype system in a one-year time frame (during the COVID lockdown) required trade-offs in the design and testing, including minimal intermediate filtering, selection of more conservative and higher-power radio frequency (RF) parts, testing with narrow-band unmodulated signals, and processing of the recorded multiplexed signals in non-real time via MATLAB. The successful demonstration of this technology calls for a spiral development effort addressing the trade-offs to realize a real-time operational system. The potential is a receiver system that operates with the properties of a conventional phased array system, that is, high-antenna gain/directivity and co-channel interferer suppression, yet one that offers these benefits simultaneously across the entire field of view of the array. This enables detection of weak and short duration signals from any direction. While this new architecture may require a narrower frequency band if the sampling rate is limited, it can convert nearly instantaneously into a conventional phased array system for collection of wide-band signals from a given direction.

Motivation: Benefits and limitations of phased array antennas

A single antenna has a fixed gain pattern with respect to the azimuthal (Az, θ) and elevation (El, φ) angles, typically where the maximum gain, the main lobe, defines the broadside direction ($\theta, \varphi=0$). The main lobe must be physically oriented toward the signal source to maximize the detection of a signal. This limitation can be mitigated by using a phased array antenna, composed of N identical antenna elements spaced at a fixed interval, as illustrated in [figure 1](#). The phase of each signal received by the antenna elements are adjusted to coherently add the gain patterns of each element in the array and steer the resultant narrower and enhanced gain pattern toward the intended signal of interest (SOI) [1]. If the direction of the SOI is not known, the main lobe may be scanned across θ and/or φ (both for a two-dimensional array) to search for the signal. In addition, if two transmitters are broadcasting at the same frequency, then steering the main lobe towards one transmitter will result in a low gain in the direction of the other transmitter, thereby reducing co-channel interference that might result in communications errors. Hence, once a set of phases are chosen to orient the array's gain in one

direction, if a SOI arrives from another direction, it will not be detected.

Smart antennas

The limitations of the conventional phased array system can be overcome through high-end digital signal processing (DSP). The signal from each antenna is fed into its own receiver whose output is digitized by an analog-to-digital converter (ADC). Each receiver and ADC together are operated under computer control and constitute a software-defined receiver/radio (SDR). If the receivers share a common reference signal, their output digital streams are coherent in phase and the system is referred to as a coherent multi-channel receiver (CMCR). These N data streams may be recorded and post-processed to implement digital phase delays to scan the high-gain composite array lobe across Az/El and to null out interferers. A system implementing a field-programmable gate array (FPGA) or another DSP architecture may be able to combine these N digital streams in real time and alter their amplitudes as well as the phase, a process referred to as digital beamforming. These systems are generally referred to as *smart antennas* or *adaptive arrays* [1]. They allow forming multiple beams to

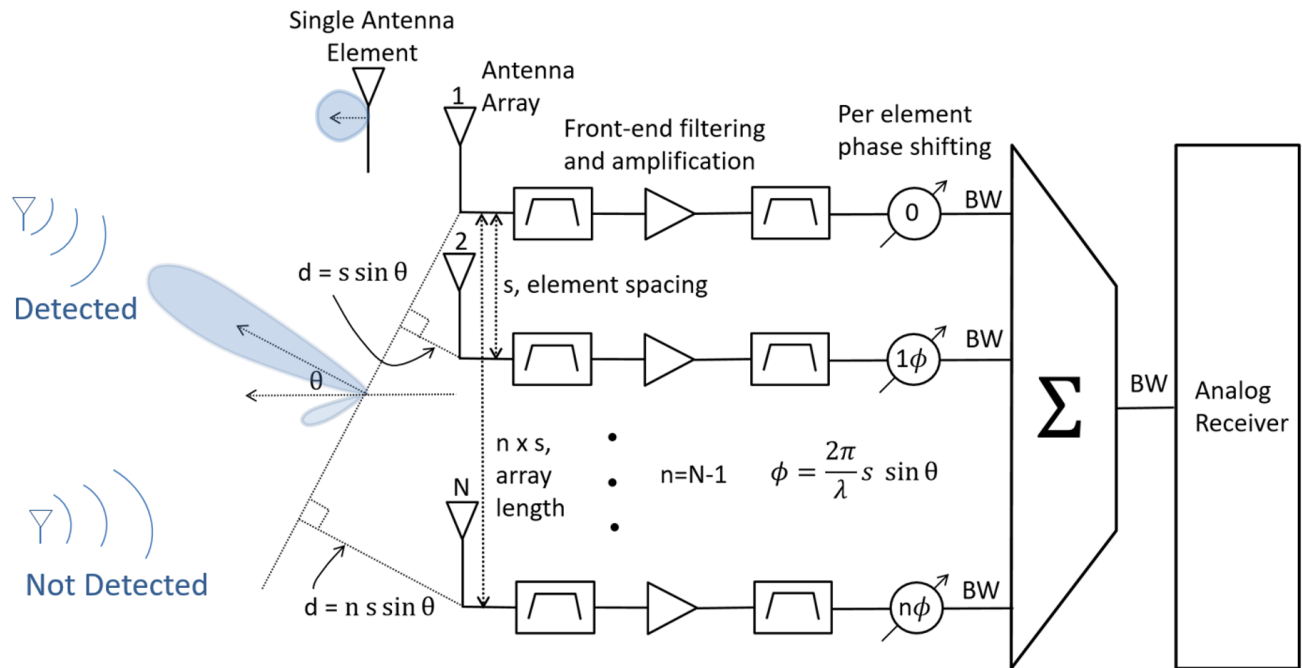


FIGURE 1. This conventional phased array antenna concept illustrates identical antenna elements spaced at a fixed interval.

simultaneously receive multiple transmitters on the same channel. The real-time algorithms they implement are commonly used in tracking mobile user equipment in a fading multipath environment and, in this implementation, are referred to as multiple input multiple output (MIMO) systems—an active area of research [2].

The frequency multiplexed phased array (FMPA) concept

Here we present a similar but novel architecture, the frequency multiplexed phased array (FMPA), as illustrated in figure 2. A signal of known bandwidth (BW) from each antenna element is converted to an intermediate frequency (IF) using an RF mixer. The mixer is a nonlinear element, common to most receivers, that combines the RF input signal with a user-supplied local oscillator (LO) tone (sine wave) to translate the signal to an IF , $f_{IF} = f_{RF} \pm f_{LO}$, where in figure 2, we select the downconverted signal $f_{IF} = f_{RF} - f_{LO}$ by filtering out the upconverted signal $f_{IF} = f_{RF} + f_{LO}$ after the mixer. However, the signals from each of the N antenna elements in the array are downconverted to different, non-overlapping and adjacent, evenly spaced IF bands, or a comb of IF s, by appropriately choosing a comb of N LO frequencies, each separated

by BW . These N LO s are coherent with each other because they are generated by a master LO , such as a shared 10 megahertz (MHz) reference signal, and each LO may add its own phase shift, ϕ_n , to correct for variations in the phase through each DC. Finally, the N downconverted signals are aggregated onto a single cable using an RF combiner. Thus, the composite signal now has a bandwidth of $N \times BW$. A single-channel SDR receives and digitizes this signal and streams both in-phase (I) and quadrature (Q) components to a computer for further digital processing or storage.

A follow-on DSP chain de-multiplexes the composite IQ waveform into the original N channels that originated from the antenna elements, and then applies appropriate phase shifts to steer the antenna gain pattern towards a target signal captured in the recording. Thus, as with the CMCR, we are beamforming on the recorded digital signal so we can look for a signal in any direction within the limits of the antenna array's field of view during the collection. Both the CMCR and the FMPA provide continuous high-gain surveillance over a large solid angle. In addition, with beamforming we suppress a second signal coming from another direction to prevent interference and can then adjust the digital beamforming again to detect and process that second signal while nulling the first. An advanced radar system employed a

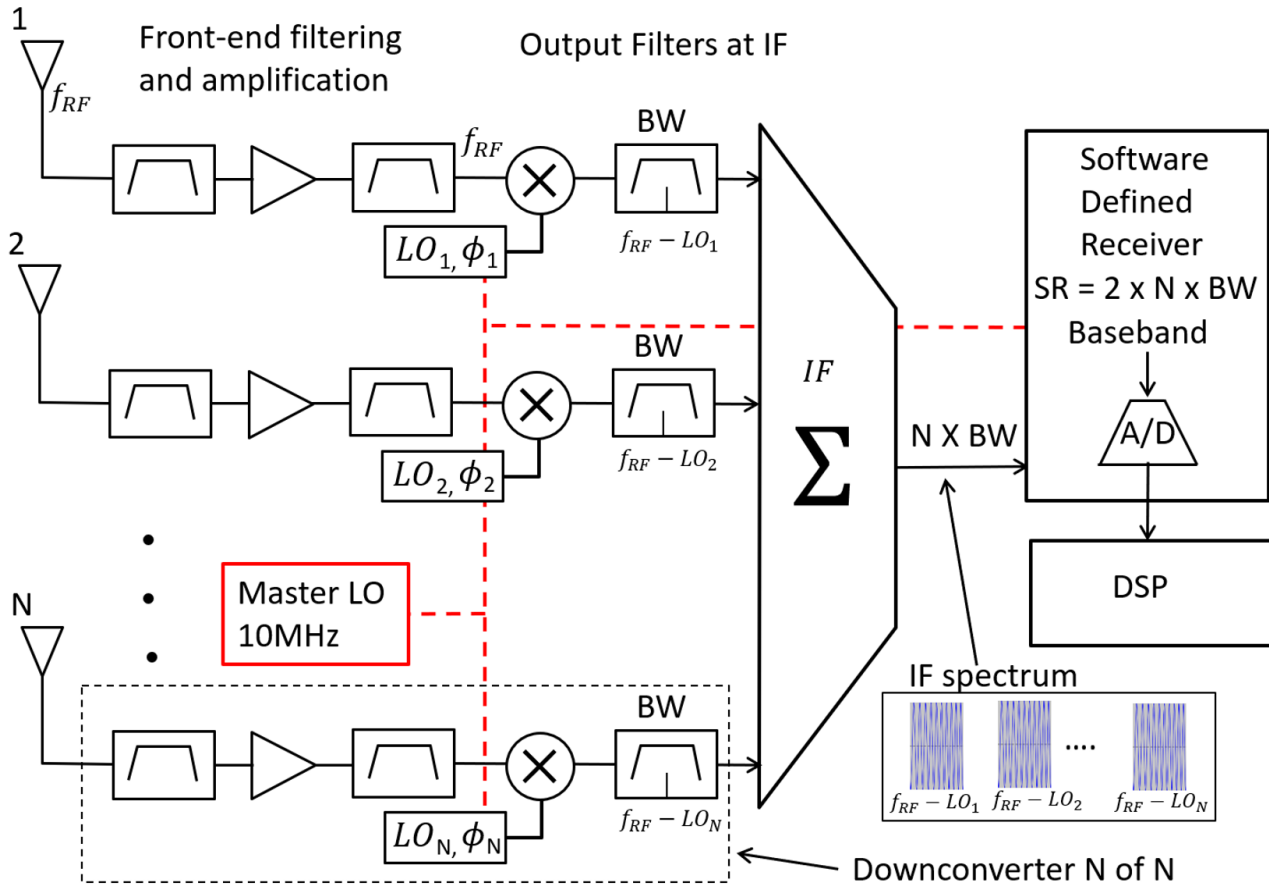


FIGURE 2. This novel frequency multiplexed phased array (FMPA) concept also overcomes the limitations of the conventional phased array system.

frequency multiplexing technique several decades ago [3]. However, that implementation continuously swept the main lobe across 180 degrees and carefully timed the sampling of the combined output to select only one beam direction. That analog approach leads to issues if the sweep rate is insufficient to adequately sample a modulated signal. Our DSP-based FMPA concept does not sweep the beam and hence is not hampered by these analog-processing issues.

Compared to a standard receiver capturing a signal of BW , the output of the FMPA is $N \times BW$, and so according to the Nyquist criteria [3], to faithfully record the signal, a sampling rate of $2 \times N \times BW$ is required compared with $2 \times BW$. Thus, the FMPA (and CMCR) records N times more data than a conventional single-channel digital receiver. If the BW of the signal and the higher data recording rate for that BW is not feasible due to a limited ADC sampling rate, an alternate concept for the FMPA involves using the FMPA

as a survey tool where it receives a narrower band at the center of the wideband signal. With near real-time processing of the lower rate signal, the FMPA system will be able to determine the direction of the target SOI. The FMPA can then collapse the frequency comb to a single frequency and simply control the phase, ϕ_n , of each element to steer the beam towards the SOI direction, allowing collection of a wider band. Thus, the FMPA system can be converted to a conventional phased array system to collect the higher BW signal. The concept of switching the FMPA system from a narrow-band surveillance mode to a wideband signal collection system is a unique capability of the FMPA system.

Other potential benefits of the FMPA over the CMCR is that the later requires N full receivers including N ADCs and must deal with the complexity of merging the N data streams. The FMPA has only one receiver, one ADC, and one data stream. Given

some number of array elements/channels, we expect this will result in a size, weight, and power (SWaP) reduction and cost savings for the FMPA over the CMCR. In addition, the FMPA DCs can be placed closer to the antenna elements and the multiplexed signal will be combined onto a single cable for conveyance to the single SDR, so the phase relationship for each channel is set near the antenna. In the CMCR, there will be separate (and potentially long) cable runs for each antenna element. The variations in length or temperature of each of these cables may make it more difficult to maintain coherency of the individual channels. However, while an N -array CMCR produces as much data as an N -array FMPA, an advantage of the CMCR is that the required sampling rate for the ADC on any of the individual SDRs in the CMCR is the same as a single channel SDR, $2 \times BW$ of the signal; whereas, the FMPA requires a sampling rate of $2 \times N \times BW$, N times higher than the CMCR. This may limit the FMPA architecture to applications with narrower band signals.

FMPA proof-of-concept development effort

A proof-of-concept (POC) FMPA-based receiver can help validate the FMPA architecture, and so the following sections describe our development and testing of an S-band [2-4 gigahertz (GHz)] FMPA prototype system. S-band not only offers many user applications [e.g., Wi-Fi; industrial, scientific, and medical (ISM) band communications; satellite communication; radar], it also yields conveniently sized antennas for development and testing. It is important to note that in any receiver, sampling rates exceeding the Nyquist rates cited above are required to address non-ideal band pass filters with gradual roll-offs. In our case, a slower filter roll-off can also necessitate larger comb spacing and an increased sampling rate. In fact, the requirement for N -distinct narrowband-pass IF output filters in [figure 2](#) required trade-offs in the development and test signals as discussed below.

To develop the FMPA prototype to capture and process signals within a 100 MHz band centered at f_c in S-band, we did the following:

1. Developed an antenna element and antenna array centered at f_c , with BW of 100 MHz. The primary array will be a 1 x 16 linear array, giving an array gain of 12 decibels (dB, i.e., 16 times that over one element).
2. Developed programmable RF hardware, a DC, to downconvert from S-band (f_c) to 500 MHz.

3. Developed firmware and a communications architecture to set the parameters of the individual DCs and control overall FMPA operation.
4. Integrated the 16-element antenna array, the 16 DCs, firmware, with power and communications distribution into the FMPA system and provided control via a (laptop) PC, including lab testing to verify operations and performance.
5. Integrated an SDR system for recording the composite signal.
6. Processed the recorded composite FMPA signal, by de-multiplexing the composite signal into the original signals received at each antenna element and then applied digital phase shifts to implement beam steering. As this is a prototype effort, processing of data was performed in non-real time using MATLAB code.
7. Performed static ground testing to validate that the FMPA exhibits the basic properties of a phased array antenna, that is, antenna directivity and nulling of interferers.

Antenna element and array development

The objective of the antenna array development was fabrication of antenna elements and the overall array with the center frequency of f_c in S-Band, each with a bandwidth of 100 MHz, and sensitive to both horizontal and vertical polarizations. We chose a printed circuit board (PCB)-based patch antenna to achieve a secondary objective of low size and weight. Because basic microstrip patch antennas tend to be narrow-band, we used a stacked-patch geometry design to increase bandwidth, whereby a single patch (near the ground plane) is reactively excited with a slot/aperture, and subsequently resonates a secondary patch that is more distant from the ground plane. Based on the size and coupling between the patches, as well as the thickness of the substrate between them, the bandwidth can be increased to cover the desired frequency range [5, 6]. We designed a feed that would create circular polarization to be responsive to both horizontal and vertical polarization.

The basic design includes two PCB layers to form the upper and the bottom patch elements, separated by a dielectric foam spacer layer (FR-3700), with a cross section, as shown in [figure 3\(a\)](#). The antenna elements are fabricated in a 1 x 4 array, with an x-ray view is shown in [figure 3\(b\)](#). Layers 1 and 2 (L1 and

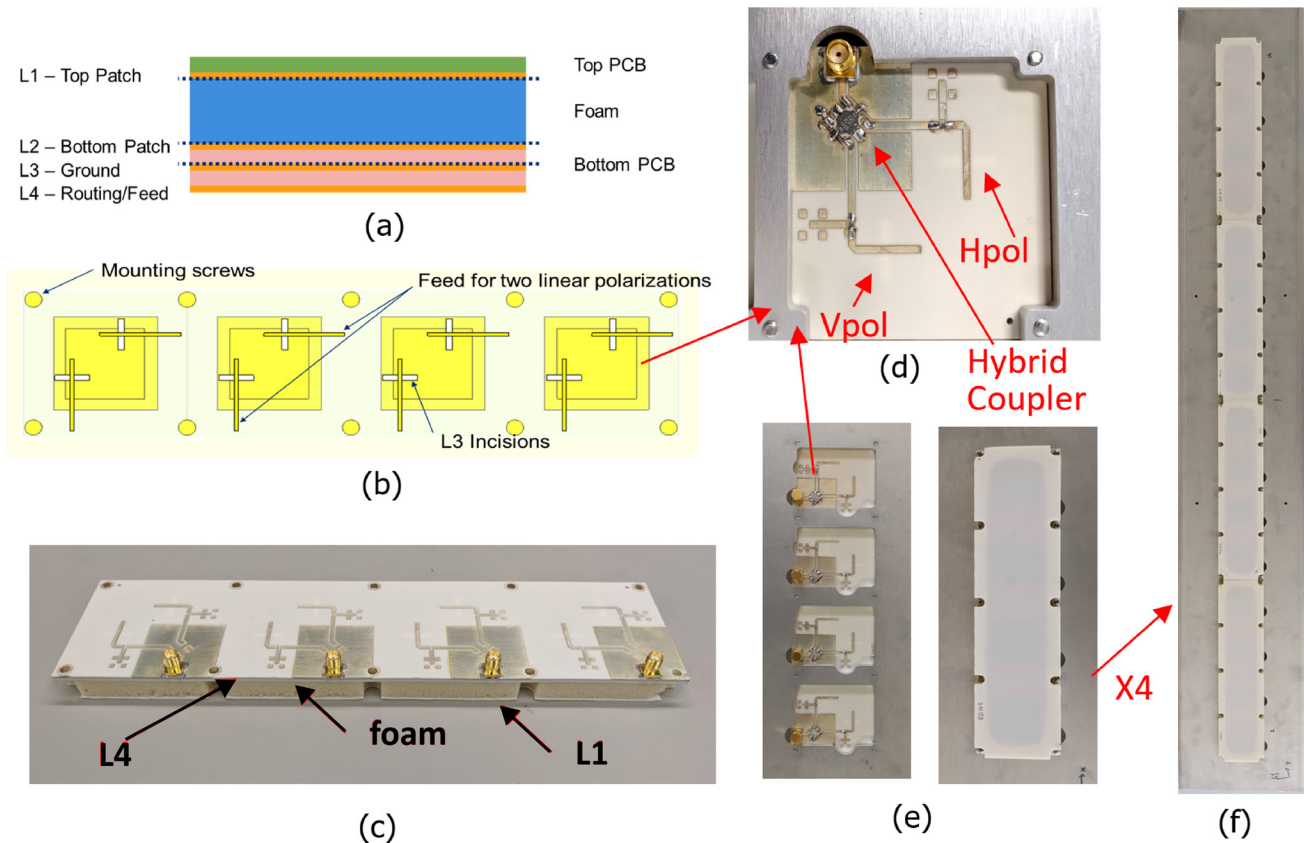


FIGURE 3. This diagram of our antenna array development shows (a) two PCBs combined with a foam spacer layer to provide the basic structure for the S-band patch. Also shown are (b) an x-ray view of four patches and feed structure, (c) a 1 x 4 building block array before gluing, (d) a hybrid coupler and feed, (e) an assembled 1 x 4 array, and (f) a 1 x 16 array.

L2) each host patches with slightly different sizes. The routing/feed layer, L4, hosts microstrip transmission lines, a SubMiniature version A (SMA) connector that injects the RF signal, and a quadrature hybrid coupler that produces a 90 degree phase shift to induce circular polarization [figure 3(c, d)]. Incisions (rectangular apertures) are made in the L3 ground layer, where the signal is reactively coupled from the feed layer to the patches, one for horizontal and one for vertical polarization. CST studio, a full wave electromagnetic solver, was used to adjust the size of the patches, apertures, and foam thickness to achieve the proper antenna center frequency and bandwidth. Fabrication of the 1 x 4 antenna array elements is summarized in figure 3(a-e), which were then combined into a 1 x 16 array [figure 3(f)].

We tested the 1 x 16 array from figure 3 to characterize its performance in an anechoic chamber. Each antenna element can influence its neighbors, and in particular, the elements near the edge of the array may exhibit anomalies as they are missing

neighboring antennas. Hence, we tested each antenna element separately while 50 ohm (Ω) terminators were connected to all the other antenna feeds. The array was rotated in the azimuth, from 0–180 degrees by roughly 5 degrees, and then the frequency is scanned from $f_c - 100$ MHz to $f_c + 100$ MHz. Each array element exhibited a gain of roughly 4–5 decibels relative to an isotropic emitter (dBi) at broadside, which also varied by 2–5 dBi over the 100 MHz frequency span. These results agree well with the CST simulations. The 3 dB beam width of a single element was found to be approximately 126 degrees, and the 16-element array factor gives a 10 dB beam width of approximately 11 degrees.

RF electronic development, downconverter design, and integration

The RF electronics for the FMPA consists of a set of $N=16$ frequency DCs that each receive a signal centered at f_{RF} , such that $f_c - 50$ MHz $< f_{RF} < f_c + 50$ MHz,

and each downconverts its signal to one of the comb frequencies centered at $IF = 500$ MHz, as shown in figure 2. We chose this center IF because we have found many SDRs, including those using the popular Analog Devices family of RF integrated circuits (e.g., Analog Devices AD9361) perform well at this frequency. Since N is even for our prototype, these downconverted bands should be centered at $f_{nIF} = (n - (N+1)/2) \Delta IF + IF$, where, $n = 1 \dots N$. In the testing described below, we chose $\Delta IF = 1$ MHz, and so our 16 comb center frequencies are at 492.5, 493.5, ..., 499.5, 500.5, ..., 507.5 MHz.

The overall block diagram for the DC is shown in figure 4. In summary, the RF signal from the patch antenna is passed through a surface acoustic wave (SAW) roofing filter centered at f_c with a bandwidth of 100 MHz to remove any out-of-band signals. This is followed by a low-noise amplifier (LNA, Qorvo TQL9093) that was chosen for its low noise figure, robustness, and high third order intercept (IP3), ensuring linearity over a large range of signal levels. The superb performance of this LNA, which is used four times in this circuit, comes at the cost of a higher required power, but we wanted to use conservative parts to prove the FMPA concept. The signal is filtered again to remove any spurious frequency

products and amplified. As the LNA has a fixed gain of 20 dB, it is combined with a programmable Peregrine (digitally adjusted) attenuator to adjust the final output so that all DCs provide the same signal level. The two other variable attenuators (combined with same LNA) in the circuit are manually adjusted with a set screw. After the signal has been boosted by 40 dB, its phase can be phase shifted in roughly 5.6 degree steps via the digitally programmable MACOM phase shifter. The amplified and shifted signal is now fed into the RF port of the mixer where it is combined with the LO tone from the phase-locked loop (PLL), to produce the downconverted IF signal $f_{IF} = f_{RF} \pm f_{LO}$. The mixer also produces a higher upconverted frequency which is removed with the image reject filter before final amplification and output to a 16-way combiner and then to the SDR for recording.

The LO frequency for the mixer is provided by the Analog Devices ADF4351 (35 MHz–4,400 MHz) PLL. This device is configured using six 32-bit registers that contain 38 digital fields. Data for the PLL, as well as one byte each for the digitally adjustable attenuator and phase shifter are sent over the serial peripheral interface (SPI) bus from a microcontroller on each DC. The PLL fields allow us to set both the frequency and the phase of the output signal relative

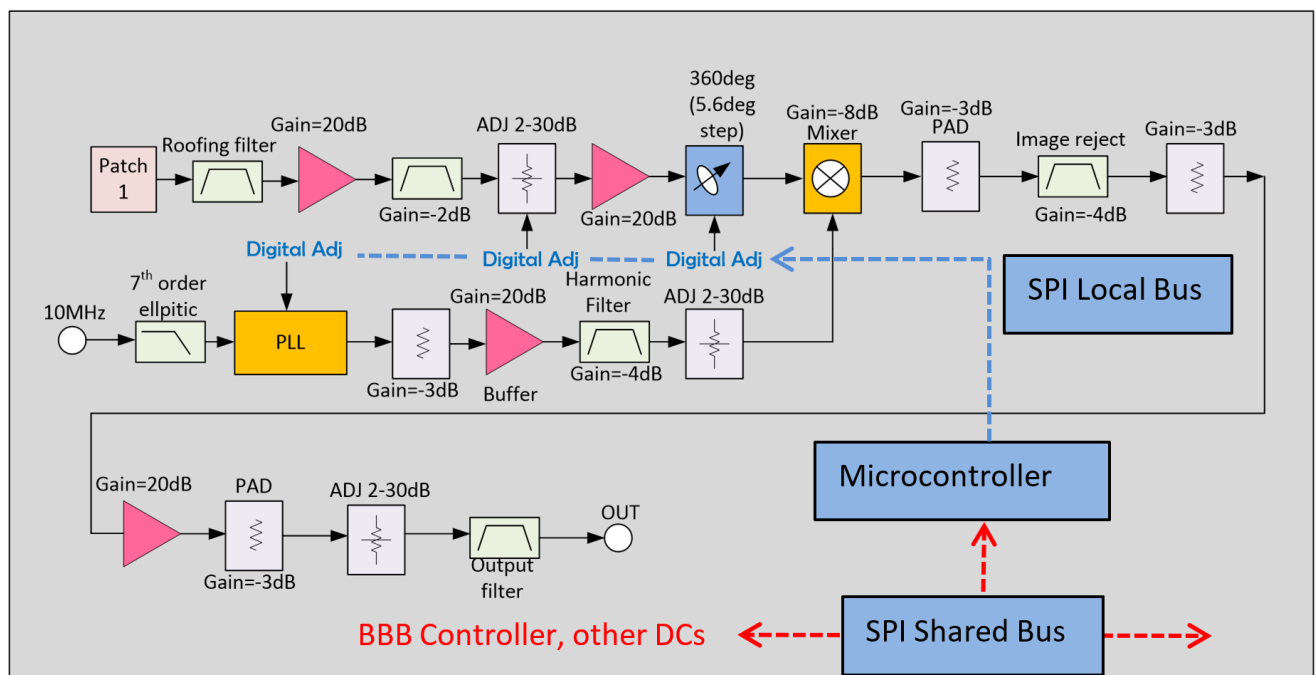


FIGURE 4. This schematic for our DC board shows filtering and amplification of the signal from a single antenna, mixing of that signal with the LO signal from the PLL to produce the downconverted signal, and finally output amplification and filtering. Communications and control are provided by a microcontroller.

to the phase of the incoming shared reference signal. The PLL provides finer control of phase, at a highest resolution of 0.1 degrees to complement the MACOM phase shifter. A major source of concern in the RF design is the introduction of stray signals from other DC boards coupling onto the 10 MHz reference signal line, which is shared among all DC boards. Since the PLL feedback loops have approximately 80 dB of gain, any stray signals can cause a proliferation of spurious output PLL signals. Hence, the seventh order elliptical low-pass filter cutting off at 100 MHz is a critical component of this design.

Trade-offs

The output filters on each downconverted channel in the comb should be centered at the different frequencies of the IF comb, per [figure 2](#). For example, if the bandwidth of the SOI and the ΔIF is 1 MHz, and the IF output of first DC is at 492.5 MHz, then there should be a band-pass filter from 492–493 MHz, 493–494 MHz for the second DC..., and 507–508 MHz for the 16th DC. If there is no noise or other signals outside of the downconverted signal's BW , then the signals in adjacent channels do not overlap. However, if there is significant noise in the channel or if the actual BW of the SOI exceeds the IF spacing in the comb, then adjacent channels in the comb will overlap and degrade the signal-to-noise ratio (SNR) of the multiplexed signal. We could develop 16 distinct 1 MHz band-pass filters, centered at each IF in the comb, using custom SAW filters, but that would require a long lead time and great expense. It would also complicate assembly of the final system with different non-interchangeable DCs, and each DC would have a fixed ΔIF and a fixed IF . An alternative solution for a follow-on development is to add another fixed IF stage in each DC where the signal is first passed through a user-selectable narrowband-pass filter in a bank of filters, with the output mixed again to the target comb frequency. We opted for the same output band-pass filter of approximately 100 MHz bandwidth on each DC for this POC system and decided that, during testing, we would limit the total BW of the signal to the IF spacing and assess the effect of channel noise on performance of the POC system.

Communications and control firmware

All 16 DCs must be set in a consistent manner to generate the comb of LO s supplied to the mixers

that, in turn, generate the required IF frequencies. As illustrated in [figure 4](#), this is accomplished via a two-tiered communications architecture with a BeagleBone Black (BBB) single board computer sending commands over a shared SPI bus as the master to 16 microcontrollers, one on each DC, which then uses a second local SPI channel to configure the three programmable RF components on each DC (the PLL, a variable attenuator, and a phase shifter). A Python program on the BBB maintains consistent data for the state of each element/DC in the array, acting as the antenna array controller (AAC). That data includes the overall tuned frequency, f_{RF} , the target center IF , and the ΔIF spacing of the comb frequencies, and from these, it calculates the registers for the three local RF components. A Silicon Labs EFM32 microcontroller on each DC acts as the antenna element (AE) controller. The BBB sends the proper register data for each EFM32 microcontroller which stores that data in memory. When the BBB has updated all EFM32s/DCs, it issues a command that directs all EFM32s to simultaneously update their local RF components.

The BBB provides an Ethernet interface, so a user may log onto the BBB from a local or networked PC to run the Python program. Alternatively, we configure the BBB to start the Python controller program on boot up and listen on the Ethernet for simple text commands that it translates into the Python functions to configure the FMPA system.

Laboratory demonstration of the FMPA system

The functionality and performance of the FMPA was verified in the lab using a signal-generated sine wave with a frequency of f_c at a power of -50 decibels relative to a milliwatt (dBm) fed into a 16-way splitter/combiner that splits the signal into 16 copies, and each is supplied to the input of one of the FMPA DCs. A function of the Python AAC class configures the FMPA to downconvert the input RF to an IF centered at 500 MHz. Another AAC function then sets the ΔIF to 1 MHz, which configures the specific LO s for each DC to generate the comb of frequencies around the center IF [see [figure 5\(a\)](#)]. Computer control is demonstrated by resetting the ΔIF to 500 kHz, as shown in [figure 5\(b\)](#).

A key capability of each DC is control of the phase shift of the signal as it propagates through the DC, which offers the ability to calibrate the system.

Input to FMPA: S-Band Tone from Signal Generator, Output of FMPA: Spectrum Analyzer

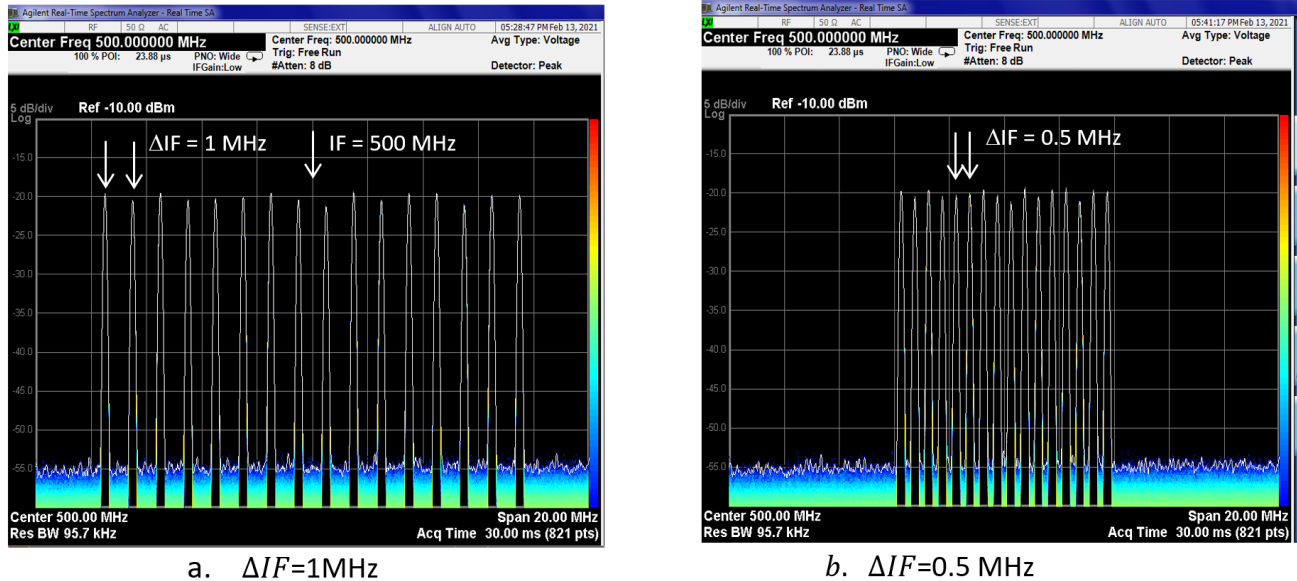


FIGURE 5. The functionality and performance of the FMPA was verified in the lab. The frequency spectrum shows successful frequency comb generation and control.

Adjusting the phase shift can also be used to operate the FMPA system as a conventional phased array as described above if the comb ΔIF is set to 0. We can use the PLL's ability to control phase to correct for the phase shift as demonstrated in the sequence of commands in figure 6, which shows sine waves from three DC outputs on a digital oscilloscope. Since ΔIF is set to 0 Hz, all three downconverted signals are at the same frequency, but due to differences in the circuits, splitters, and cables, etc., they are out of phase. The figure shows the results of Python commands executed to advance the phase of AE 2 by 330 degrees and AE 3 by 30 degrees, so that they may be brought into phase alignment with the signal from AE 1. This phase alignment is stable over time—even after shutdown and reloading these settings the next day—demonstrating the coherency of the system.

SDR collection system

The 16-channel frequency-multiplexed and combined analog output of the RF Electronics feeds an Ettus E-310 SDR. This SDR executes an embedded Linux operating system on an application processor and operates with an FPGA to collect RF samples and save them locally on the microSD card. We developed a MATLAB graphical user interface (GUI) to interface

with the RF Electronics controller and the Ettus SDR. Data is captured in one second blocks at a sampling rate of 20 mega samples per second on the SDR and then automatically extracted and saved in an RF Recordings folder on a laptop. The GUI displays a frequency domain power spectrum plot as well as a time-domain RF amplitude plot to verify that the expected target signals are on air, and for awareness of any potential interference signals.

FMPA processing

The MATLAB FMPA algorithms that process the recorded samples are illustrated in figure 7. The first step is to extract the original AE signals from the multiplexed signal by multiplying the composite waveform by the complex conjugate of the baseband comb frequency resulting from the SDR tuning and digitizing, [e.g., $f_{BBk} = (k - (N + 1) / 2) \Delta IF$], for even N and $k = 1, \dots, N$. The signal from that channel is now essentially shifted to $0 - BW$ in frequency at baseband, and signal components from adjacent channels are removed with a low-pass digital filter. As each data stream is now limited to a bandwidth of BW , they should be decimated by a factor of N to reduce the amount of data processed when beamforming. We now have a set of N distinct complex waveforms

All AE PLL phase are 0

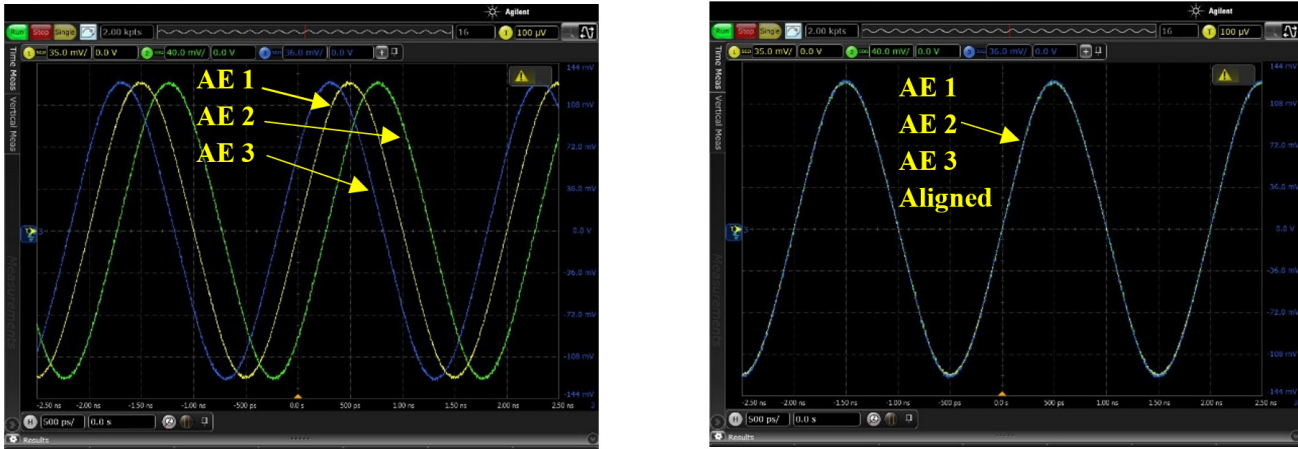


FIGURE 6. The plot on the left shows three DCs converting an input S-band sine wave to an output sine wave at 500 MHz, but with phase differences due to physical anomalies of each board. In the right plot, commands are sent to advance the phase of DC 2 and 3 respectively to bring the output of all three AEs/DCs into phase alignment.

corresponding to signals received at the individual AEs. In order to beamform these signals and look in a specific angular direction θ with respect to the broadside of the antenna array, we multiply each by a complex phase coefficient. The array of coefficients is determined by the calculated phase shift, $\phi = 2\pi/\lambda s \sin \theta$, where s is the spacing of the antenna array elements, λ is the wavelength of the carrier (i.e., $\lambda = c/f$), and c is the speed of light. Thus, the individual phase coefficients are $e^{jk\phi}$, where $k=0, \dots, N-1$, corresponds to the individual channel and antenna element. However, we must compensate for the individual phase shifts through each DC, (ϕ_{ck}), as shown in figure 7. If those phase corrections are not programmed into each DC, then they must be accounted for in the processing. Finally, the individual channels are added together to form the total beamformed signal at a given azimuthal direction (θ). The FMPA is now pointing its high-gain lobe at the target azimuth angle while providing rejection from interference signals and targets at other azimuthal angles.

FMPA static outdoor range test

The FMPA POC system with its 1 x 16 linear array was tested at a local outdoor RF test range. The comb spacing was set to Δf of 1 MHz, as shown in figure 5. Two signal generators, TX1 and TX2, with Vivaldi horn antennas, were employed as target transmitters, broadcasting a continuous wave (CW) tone at $f_c + 50$ kHz for TX1 and f_c for TX2. The different frequencies

of TX1 and TX2 enabled convenient spectral separation of the two targets for comparison of beamforming rejection. Placement for the FMPA, TX1 and TX2, are shown in figure 8 and in the accompanying table. Each transmitter was also moved to an alternate position during the testing (TX1-Alt, TX2-Alt). Phase and amplitude calibration of the individual antenna channels was performed using the known frequency and position of one of the transmitters to determine ϕ_{ck} in figure 7. Amplitude correction (not represented in figure 7) was necessary because the main lobes of TX1 and TX2 were not pointed accurately at the FMPA system and were directed above the FMPA to reduce multipath reflections.

The objective of this test and POC demonstration was to verify that the FMPA system can steer its antenna lobe towards the desired target while rejecting energy from the other target. This was achieved by numerically forming a beam and steering it in 128 steps from $-\pi/2$ to $+\pi/2$ azimuth angle while recording the total power received in the beam at each step. The resulting swept beam power profile versus azimuth is then compared to the theoretical power profile from the known position of the emitters. The results before ($\phi_{ck} = 0, k=1$ to N) and after calibration ($\phi_{ck} \neq 0$) are shown in figure 9. The blue trace represents the actual beamformed result, while the gray trace represents the predicted response based upon the known location of the targets. When the phase and amplitude calibrations are applied, we have

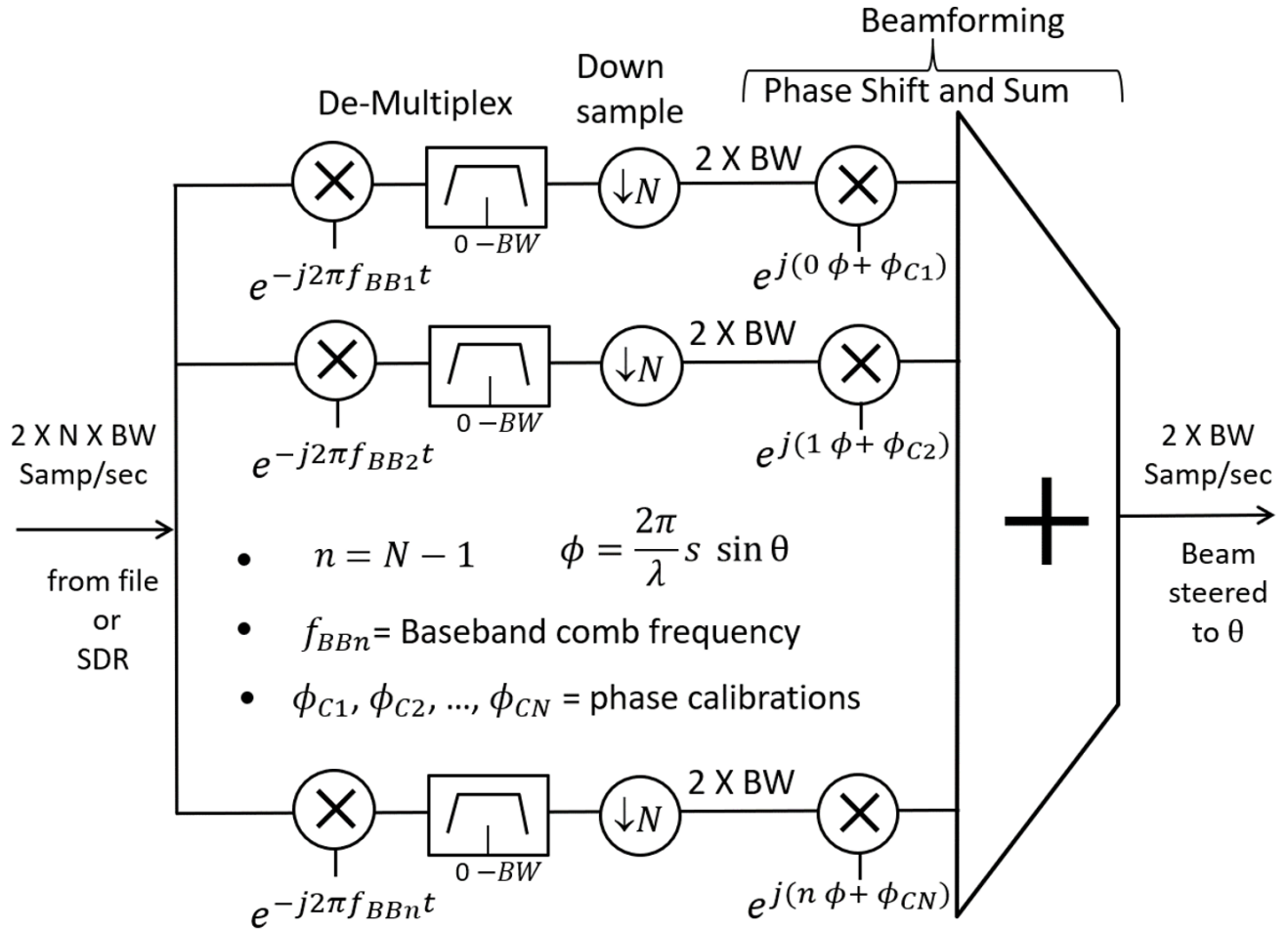


FIGURE 7. This diagram illustrated the FMPA signal processing on the digitized multiplexed signal from the SDR in order to steer it to receive a signal arriving from an angle θ .

very close agreement between the theoretical and measured profiles [figure 9(b)]. Our testing results showed good agreement between the theoretical and experimental results independent of whether we chose TX1 or TX2 as the calibration source.

Since the frequency of the TX1 tone was 50 kHz higher than TX2, but still well within the 1 MHz signal bandwidth of the FMPA comb spacing, it was possible to compare the individual signal strengths of the two targets as received after beamforming on TX1 and after beamforming on TX2. This comparison is an effective means for determining the rejection of the off-beam target. The antenna responses predicted by theory for beams looking at TX1 and TX2 are shown in figure 10.

When the beam is pointed at TX1, the measured rejection of a signal emanating from TX2 is approximately 25.7 dB compared to the predicted rejection of 33 dB [figure 10(a)]. For the beam pointing at TX2, the rejection of a signal emanating from TX1 is approximately 35.6 dB, compared with the predicted value of 32 dB [figure 10(b)]. These results demonstrate that the FMPA processing can reject co-channel interferers.

We moved TX1 and TX2 to their alternate azimuthal locations on the range, (TX1, TX2-Alt) and (TX1-Alt, TX2-Alt), to provide greater confidence in the beamforming success above. The close agreement between predicted and measure beamformed intensity across azimuth for these two additional

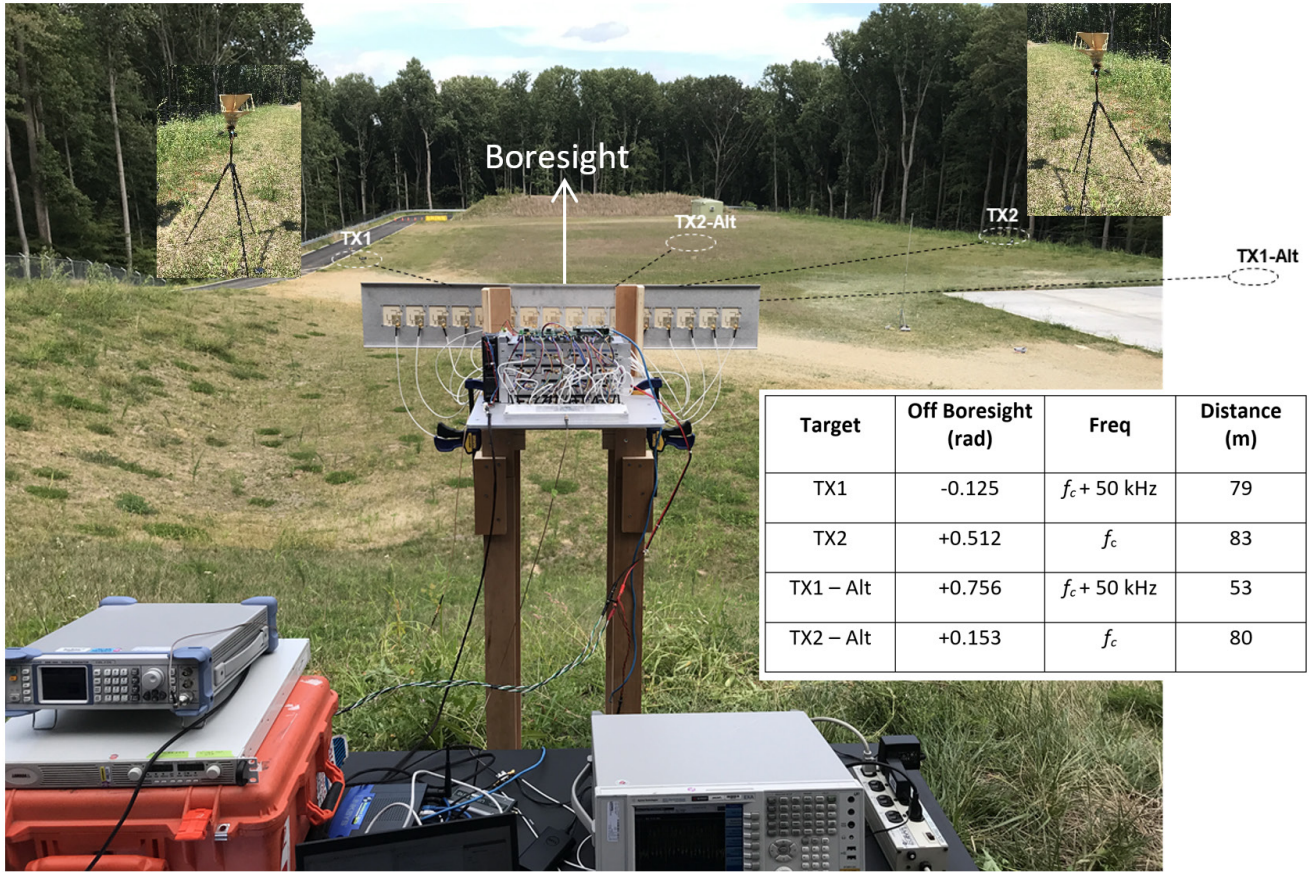


FIGURE 8. This photo and markup show the placement and configurations of the FMPA system and two target transmitters (TX1 and TX2) on test range.

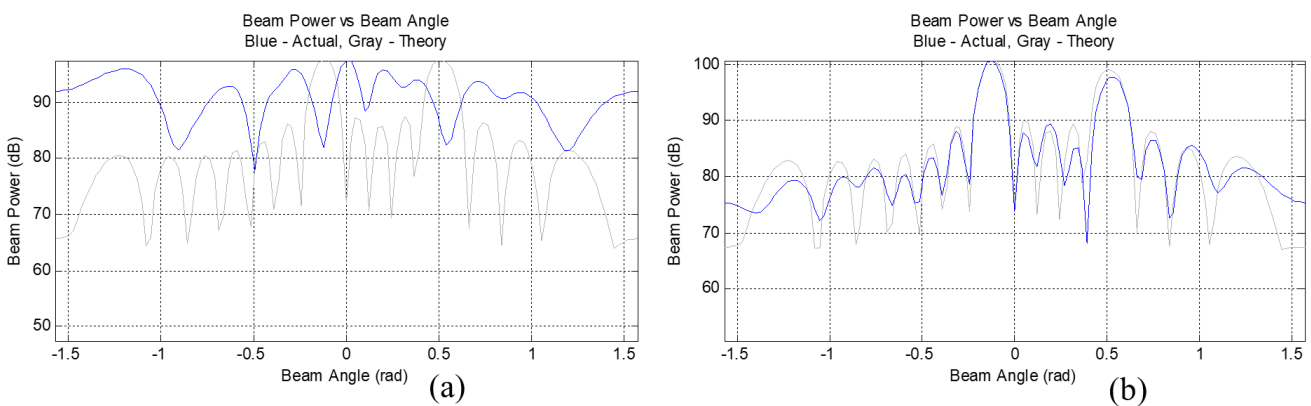


FIGURE 9. These plots show the experimental swept beam power profile (blue) and the theoretical profile (gray), (a) before calibration, and (b) after calibration.

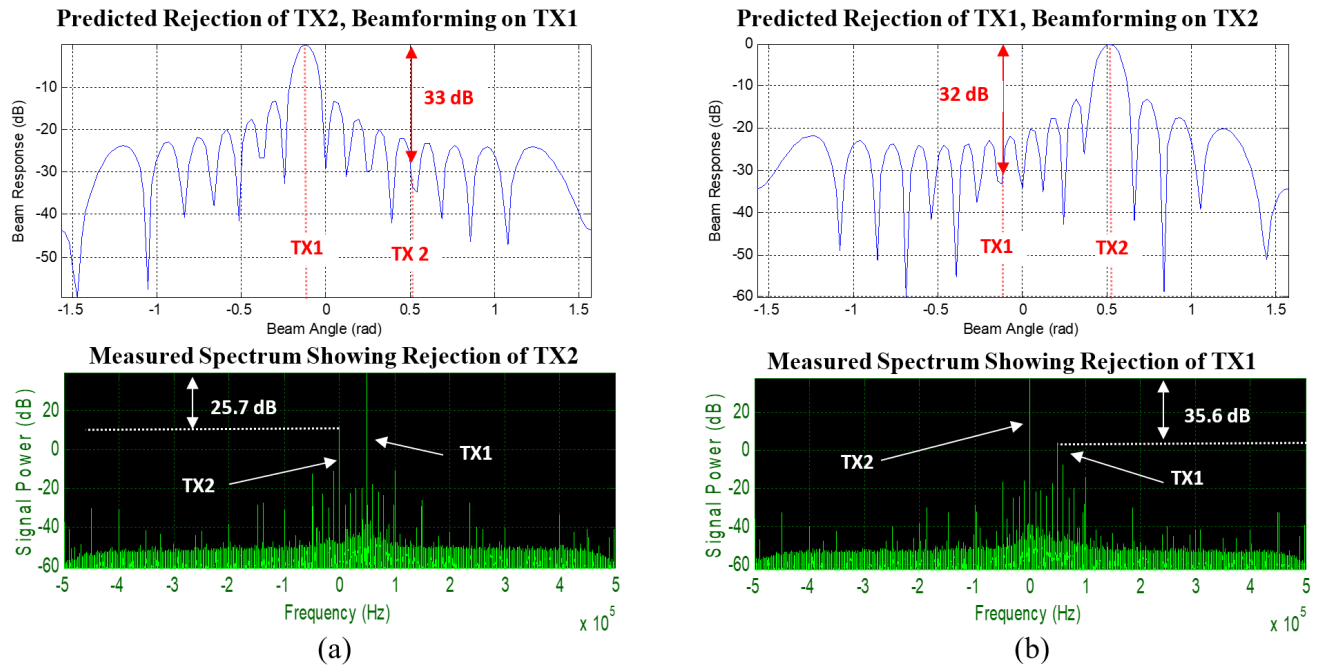


FIGURE 10. This measured data swept beam power profiles (top) along with corresponding spectral data (bottom) to demonstrates co-channel interferer suppression of (a) TX2 while beamforming on TX1, and (b) TX1 while beamforming on TX2. Bottom plots are spectra centered at f_c .

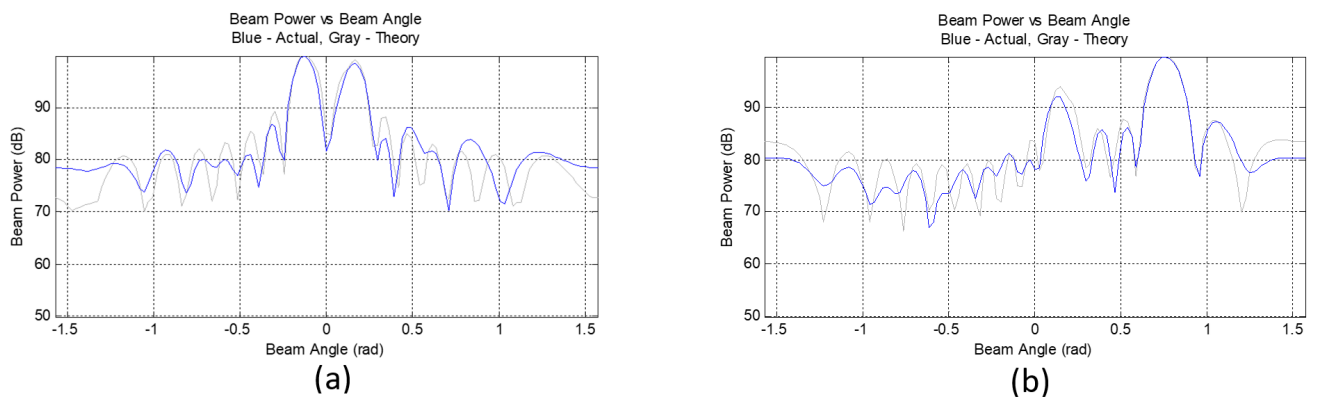



FIGURE 11. This data shows experimental (blue) and theoretical (gray) swept beam power profiles after moving TX1 and/or TX2 to alternate locations on the test range, with specific configurations of (a) TX1, TX2-Alt, and TX1 (b) TX1-Alt, TX2-Alt.

configurations, as shown in [figure 11](#), demonstrates convincingly that the FMPA system is offering the capabilities of a conventional phased array through digital beamforming.

Conclusions and next steps

This work developed a prototype and validated a design and processing implementation for FMPA systems. The FMPA approach offers the phased array benefits of a) gain pattern directivity, and b) co-channel interference reduction. In addition, c) we proved that through DSP on the recorded signal, we could steer the beam in any direction within the array's field of view. Thus, an FMPA system would not miss a weak or short duration SOI from any direction. Finally, the FMPA offers potential benefits of reduced SWaP and lower processing overhead compared with CMCR systems.

Follow-on development efforts should focus on three objectives:

1. The system should implement a user-selectable filter bank along with a two-level mixing scheme to prevent overlapping noise from adjacent combs reducing the overall SNR of the received signal.
2. Reduce the system SWaP. This POC system used a conservative LNA that required a lot of power. A lower power alternative should be implemented, which would also reduce the heat generated and the requirement for cooling fans along with heavy and bulky heat sinks.
3. The MATLAB algorithms should be converted to real-time digital DSP implementation using an FPGA. This would enable the FMPA system to convert from a narrowband multidirectional system with a wide angular view to a wider-band conventional phased array system nearly instantaneously. 

References

- [1] Balanis CA. "Chapter 6: Arrays" and "Chapter 16: Smart Antennas." *Antenna Theory, Analysis and Design, Third Ed.* Hoboken (NJ): Wiley & Sons, Inc.; 2005. ISBN: 0-471-66782-X.
- [2] Malkowsky S, Vieira J, Liu L, Harris P, Nieman K, Kundargi N, Wong IC, Tufvesson F, Öwall V, Edfors O. "The world's first real-time testbed for massive MIMO: Design, implementation, and validation." *IEEE Access.* 2017;5:9073–9088 doi: 10.1109/ACCESS.2017.2705561.
- [3] Johnson MA. "Phased-array beam steering by multiplex sampling." *Proceedings of the IEEE.* 1968;56(11):1801–1811. doi: 10.1109/PROC.1968.6754.
- [4] Proakis JG, Manolakis DG. *Digital Signal Processing, Principles, Algorithms, And Applications, Fourth Ed.* Upper Saddle River (NJ): Pearson Prentice Hall, 2007. ISBN: 0-13-187374-1.
- [5] Targonski SD, Pozar DM. "Design of wideband circularly polarized aperture-coupled microstrip antennas." *IEEE Transactions on Antennas and Propagation.* 1993;41(2):214–220. doi: 10.1109/8.214613.
- [6] Gao S, Li LW, Leong MS, Yeo TS. "A broad-band dual-polarized microstrip patch antenna with aperture coupling." *IEEE Transactions on Antennas and Propagation.* 2003;51(4):898–900. doi: 10.1109/TAP.2003.811080.

Additive Manufacturing of Electronic Circuits for Novel Applications

Daniel R. Hines

The next generation of electronic circuits will most likely not be flat, rectangular printed circuit boards (PCBs) as we are familiar with inside many of our computers and electronic gadgets. Such a form factor may be acceptable for controlling big, boxy electronics but not for sensors fitting within, say, a football player's mouth guard or helmet, or on the skin of a premature infant in a neonatal unit. What if electronic sensors could be fabricated to be flexible, stretchable, or even built right into the gadget that they are designed to work with [1, 2]? For example, it takes a lot of time and effort to take an airplane out of service for a few days in order to inspect the air frame for wear-and-tear, material fatigue, microcracks, and other lifetime aging. What if strain sensors could be fabricated right into the airplane's fuselage or wings and monitored in real-time over the life of the airplane? With such a data set, recording the history of a specific plane (or any mechanical system for that matter) could provide a very advanced understanding of the need for maintenance or for an assessment of a safe, functional lifetime of the plane. Couple this with artificial intelligence (AI) and machine learning, and an industry could create a very powerful and much safer means of understanding the integrity and lifetime of many mechanical systems, not just airplanes.



Additive manufacturing printing methods

So, how can the fabrication of such next-generation sensors and electronic circuits be achieved? Let's consider the advancements that are being made in the area of additive manufacturing. We are all familiar with three-dimensional (3D) printers, where a filament passes through a heated nozzle and is printed layer-by-layer to fabricate some mechanical part of interest. Actually, such 3D printers come in many varieties which can typically print parts out of plastic and metal materials [3, 4, 5]. There is also a subcategory of 3D printers referred to as direct-write printers which encompass syringe, inkjet, and aerosol-jet (AJ) printing [6, 7]. An example of syringe printing could be the use of a piping bag for cake decorating, while an example of inkjet printing could be an inkjet printer used to print black and white or color copies of a paper document, and then an example of AJ printing could be a spray paint system or an artist's airbrush used to paint car bodies. For additive manufacturing, utilizing such direct-write printing methods, the passive (only conveying color or optical contrast) inks in the examples above would be replaced with active materials such as metal nanoparticle inks for printing conducting features or polymer inks for printing dielectric/insulating features [8]. Equipped with such functional inks, a direct-write printer could be used to print alternating layers of patterned conducting features separated by printed dielectric layers to fabricate circuitization (i.e., wiring) layers onto a given surface that would function in a manner equivalent to the copper/flame retardant 4 (FR4) layers in a PCB. Furthermore, other functional inks having resistive, magnetic, ceramic, etc. properties could also be used in such direct-write printers to print sensor elements.

While all three direct-write printing methods can and have been used to fabricate electronic components [9, 10], there are application-specific advantages to one method over another. For example, syringe printing typically requires the end of the printing tip to track the print surface within a distance equal to half the tip diameter. For fine feature printing, this could mean tracking a non-flat surface within 10–25 micrometers (μm). This can be a daunting task for non-ideal, non-flat surfaces. Inkjet printers are typically equipped with an array of microprint nozzles configured in a straight line. This multinozzle print head typically needs to track the surface at a distance of 2 millimeters (mm) above the print surface.

Therefore, printing onto non-flat surfaces can be problematic with such a print system. When dealing with non-flat, 3D surfaces, AJ printing can offer a specific advantage over these two other printing methods in that an AJ print nozzle is set to track 3–5 mm above the print surface and therefore is rather insensitive to surface roughness and can be easily manipulated to print onto a 3D surface. For these reasons, the main body of work related to the application of additive manufacturing methods to the fabrication of high-quality electronic circuits and sensors presented below will focus on AJ printing.

Aerosol-jet printing

Currently available AJ printing tools come in two types, one where the aerosol is created using ultrasonic energy and one where the aerosol is created pneumatically. For the ultrasonic method, ultrasonic energy is transferred into an ink container such that a surface wave is created at the top surface of the ink, causing small droplets of ink to be “ripped off” the ink surface, thus creating an aerosol mist above the ink. This aerosol is then transported by a gas flow that carries this aerosol mist into a mist tube, thus creating an ink stream [11]. For the pneumatic method, much like in a spray paint can, ink is sucked up into a tube and forced through a pin hole by a gas flow stream. This Venturi effect creates an aerosol mist in the ink jar that is carried into the mist tube by the gas flow. Unlike the spray paint can, however, some of the gas flow needed to create the aerosol must be removed in order to establish a controllable ink stream, and the ink stream needs to be collimated so that it is confined to a diameter somewhere, typically, in the range of 10–200 μm . This can be accomplished by adding an aerodynamic-focusing insert and an exhaust in order to both collimate and reduce the gas flow rate of the ink stream as it enters the mist tube. A schematic drawing showing the details of an AJ print nozzle and a picture of a commercially available AJ printer printing a silver (Ag) nanoparticle ink onto a 4-inch hemisphere is shown in [figure 1](#).

Measuring ink stream dynamics

With an ink stream having been created for a given ink on an AJ printer, the volume of ink printed must be set and/or measured in order to print a feature of a specified geometry [12, 13]. This can be done by mounting an inkwell array, similar to what is shown

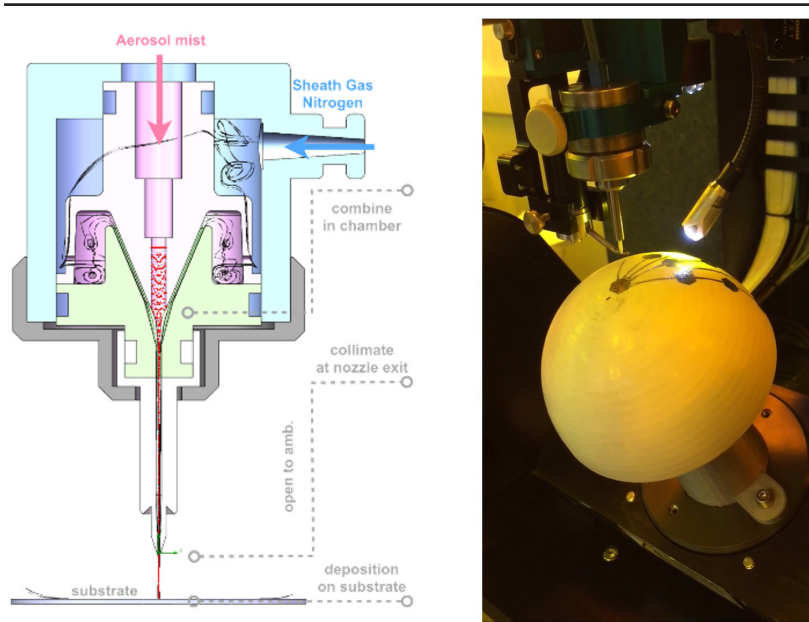


FIGURE 1. The cutaway drawing (left) highlights the ink stream dynamics within an aerosol jet (AJ) printer nozzle [11]. In the photograph (right), an AJ printer nozzle is being used to print a silver (Ag) nanoparticle ink onto the surface of a 4-inch hemisphere.

exact gas flow rates, changes in the ink over time, room temperature, and humidity, etc. Currently, there is no good way to track these changes in the ink stream, and so it is an interesting area for further research efforts [16]. Currently, the best method is to set a specific R_{ink} and then print a test trace. After post-processing, a post-processing deposition rate R_{pp} can be calculated by measuring the cross-sectional area (CS) of the test trace and multiplying that by the print speed (s) used to print the trace, such that $R_{pp} = CS * s$. At this point, the solids fraction of the ink stream used to print the test trace can be represented as a scale factor (f) where $f = R_{pp}/R_{ink}$. Using these AJ printing techniques, it is possible to fabricate high-quality electronic components within an acceptable tolerance [17].

in figure 2, onto the build plate of the printer and sequentially printing into individual inkwells of a known volume ($V_{inkwell}$) for a specified time interval ($t_{inkwell}$) and adjusting the gas flow rates until each inkwell is just filled [14].

With this inkwell method, a specific ink stream deposition rate can be established where $R_{ink} = V_{inkwell} / t_{inkwell}$. Knowing the exact deposition rate then allows for a specific volume of ink to be printed as required to print a feature with a specific designed volume, V_{design} . However, depending on the properties of a given ink, R_{ink} may not be the deposition rate that corresponds to the volume of a designed feature. This is because an ink can contain solvents, binders, etc. that are removed from the printed feature during post processing (e.g., curing, sintering), leaving only the “solids” as part of the final printed feature. Therefore, the “solids fraction” of an ink stream needs to be measured for a given ink on a given AJ printer [15]. Furthermore, depending on the dynamics of the ink stream, the solids fraction can vary depending on the

Printed hybrid electronics

What does it mean to additively fabricate an electronic sensor or circuit [18, 19, 20, 21]? Electronic circuits typically contain passive components (such as resistors, capacitors, and inductors) and active components [such as integrated circuits (ICs) wire-bonded into packages] soldered onto a PCB. Examples of printed resistors, capacitors, and inductors are

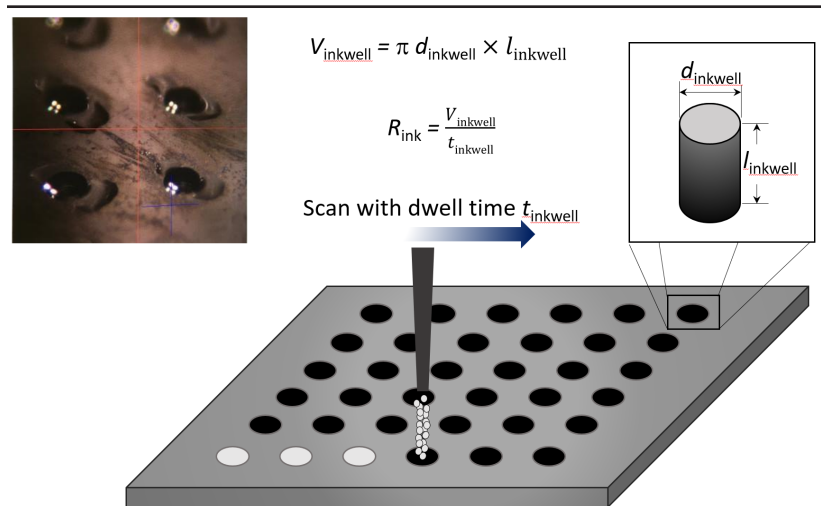


FIGURE 2. The inkwell method depicted here allows for the determination of an ink stream deposition rate R_{ink} [14].

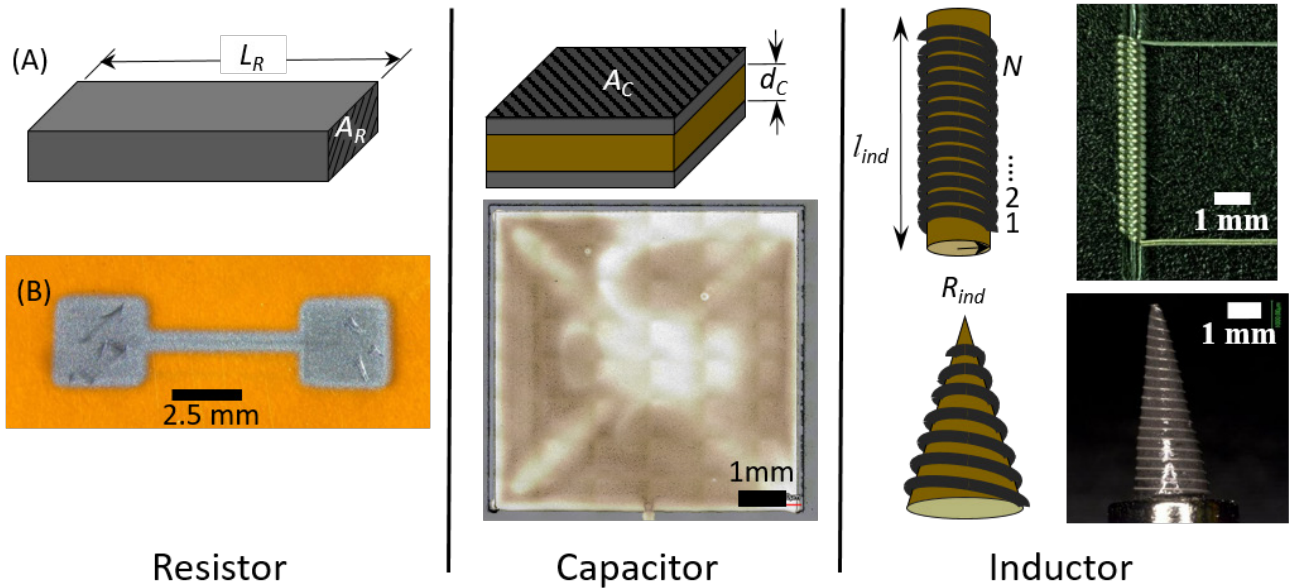


FIGURE 3. Passive components of electronic circuits can be additively fabricated as illustrated here in a printed resistor, a printed capacitor (center), and printed inductors (right).

shown in [figure 3](#); however, most ICs are too complex to be additively fabricated. For example, it is not possible to print an integrated circuit on par with an Intel 16-bit 8088 from the late 1970's, let alone a 32-bit Pentium or 64-bit Core i7 processor from the last two decades.

Nevertheless, such an IC chip can be removed from its package and used stand-alone, where the package and lead frame are eliminated and the wire bonds are replaced by printed interconnects. An example of a packaged IC is compared to a bare die with printed interconnects in [figure 4](#). This allows for a hybrid circuit approach to be developed, where components can be printed where possible and placed as bare die when printed versions are not possible. In addition to the printed and hybrid components, the PCB itself can be replaced with printed circuitization traces. Largely, it is this ability to print a replacement for the PCB that enables a variety of possibilities from rapid prototyping of circuits, to partially printed circuits, to fully printed hybrid electronic (PHE) circuits. Examples of each of these will be presented and discussed in the following section.

From rapid prototyping to PHE circuits

In [figure 5](#), two commercially available circuit boards are shown, the first one is a Mini Circuits, Model

ZFL-1000LN+, low noise amplifier (LNA) and the second is an Arduino Mini.

Both of these circuits can be modified such that rapid prototype and PHE versions can be fabricated using AJ printing methods. Let's first consider the LNA circuit in order to illustrate how additive manufacturing can be used for the rapid prototyping of

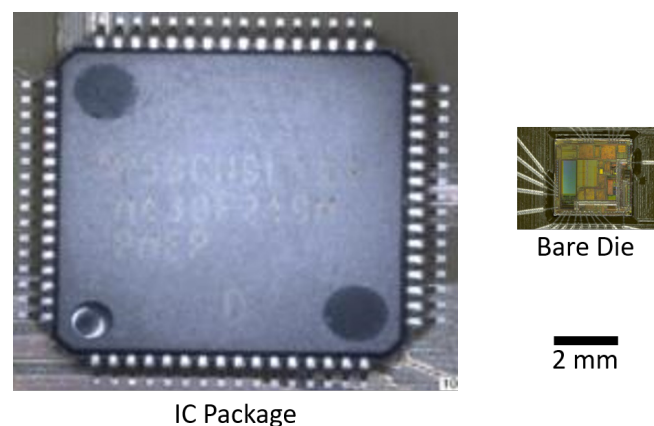


FIGURE 4. Most integrated circuits (ICs) are too complex to be additively fabricated but can be removed from their package and used stand-alone. Here is an example of a packaged IC containing the bare die microcontroller IC chip that is shown (left) as a stand-alone bare die with printed interconnects (right; scale bar related to both images).

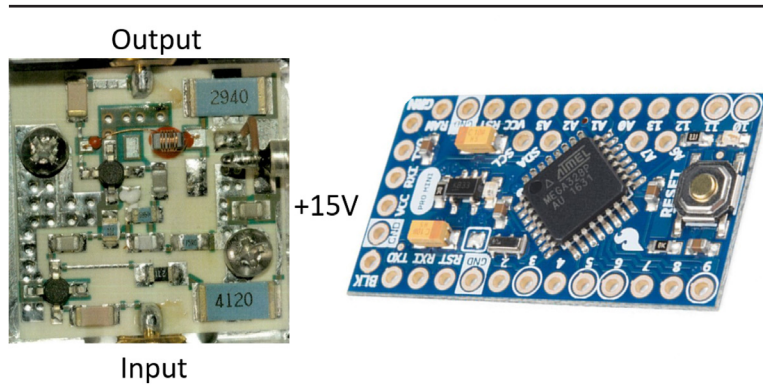


FIGURE 5. Standard commercially available PCBs—(left) Mini Circuits, Model ZFL-1000LN+, low-noise amplifier (LNA) and (right) an Arduino Mini—can be modified such that rapid prototype and PHE versions can be fabricated using AJ printing methods.

an electronic circuit. Suppose we wanted a similar circuit in a different form factor (geometry), that is, not a square geometry but rather a version that is long and skinny. Figure 6(a) shows the commercially available LNA circuit. The circuitization layout can be redesigned for a different form factor and turned into an AJ tool path that can be printed onto basically any surface. In figure 6(b), (c), and (d), versions of this circuit with circuitization are printed in ratios of 1:1 (b), 3:1 (c) and 5:1 (d) are shown [22].

Once a new design layout exists, depending on the complexity of the circuit, a new prototype can be printed in a matter of hours. At this prototype stage, the electronic components are still fully surface

mounted. One of the challenges with this is that it is not easy to solder to printed Ag (the standard AJ conduction ink), and as such, the components are typically glued in place with a dot of electronic adhesive and then electrically connected by syringe printing an Ag paste that bridges between the component and the printed trace. This method works reasonably well but is not always as robust, as many of the Ag pastes end up creating a brittle electrical connection. This is an area that provides opportunities for further research into the ability to incorporate soldering methods into additively manufactured circuitization. As additive manufacturing of electronic circuits progresses

from the rapid prototype capability to a fully fabricated PHE version of a circuit, it is possible to mix and match standard surface-mounted components, bare die versions of components, and printed components all together in a single circuit. Indeed, there will be many occasions where bare die versions of a packaged component are not available. One workaround to the soldering problem related to integrating such packaged components into an additively fabricated circuit is to use a leadless chip carrier (LCC) version that is mounted upside down in a cavity with printed interconnects. In this same manner, standard passive components can also be used prior to being swapped out for printed versions. Figure 7 shows an LCC packaged accelerometer and standard surface-mounted resistor both mounted in respective cavities.

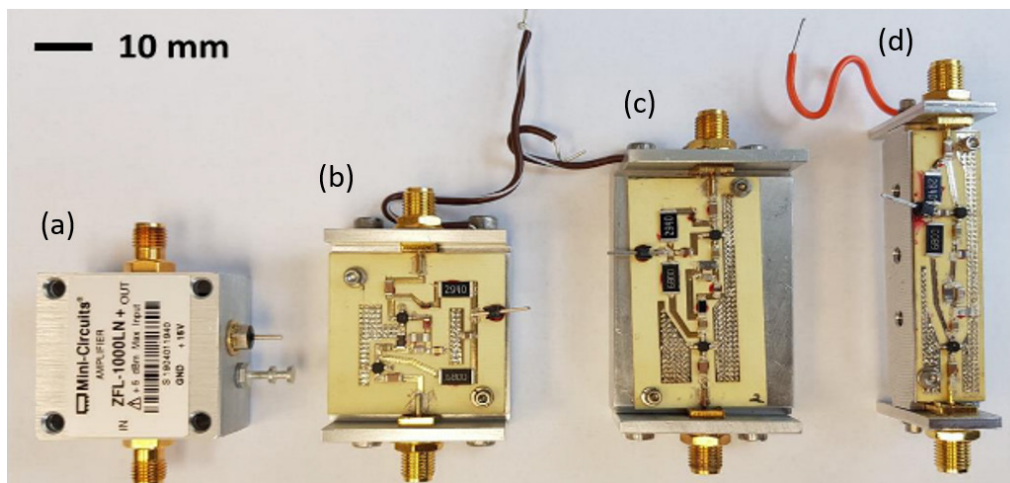


FIGURE 6. (a) For this unaltered LNA circuit, the circuitization layout can be redesigned for different form factors and turned into an AJ tool path that can be printed onto basically any surface, as seen in rapid prototypes (b) with a 1:1 ratio, (c) with a 3:1 ratio, and (d) with a 5:1 ratio.

Note that the cavity always has to be bigger than the component which necessitates a printed moat fill (red adhesive for the resistor and clear polymer for the accelerometer) to create a continuous, smooth surface onto which the interconnects can be printed. Where bare die are available, the bare die itself can be mounted onto a surface and

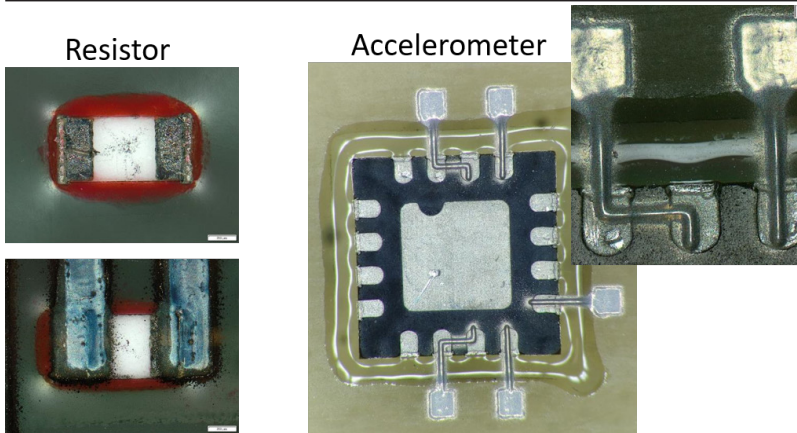


FIGURE 7. On the passive resistor (left) and packaged accelerometer (right), electronic components are mounted in cavities with moat fills printed to create a smooth transition for printed interconnects.

interconnects printed such that the electrical pads of the die are properly connected to the printed circuitization and thus, as such, properly connected into the electric circuit. [Figure 8](#) shows an example of a bare die with printed interconnects.

Just as with cavity-mounted components, the printed interconnects here also need to have a smooth, continuous surface over which they are printed. A typical bare die can have a thickness of 50–500 μm and so a “ramp” needs to be printed along the edge of the bare die in order to establish the required smooth surface [23]. Such fillets can be seen along the die edges, where needed, in the image shown in [figure 8](#). With the capabilities highlighted in [figures 6, 7, and 8](#), we can redesign the Arduino mini circuit shown in [figure 4](#) so that the circuitization for a similar PHE circuit can be additively fabricated. [Figure 9\(a\)](#) illustrates what is referred to as a three-layer circuit that represents an AJ printable, PHE circuit designed to have similar functionality to an Arduino Mini type circuit.

The red, green, and purple features map out the three circuitization layers, and the magenta features map out the component interconnects. This PHE circuit-level demonstrator contains: 1) a bare die version of an Atmega328P microcontroller

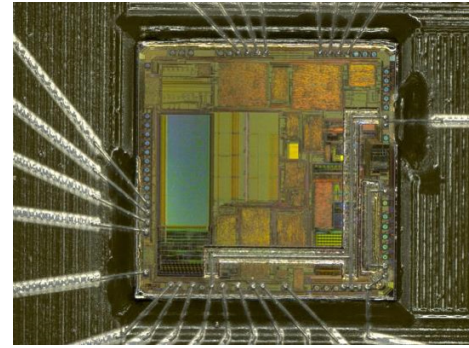


FIGURE 8. In this optical image of a bare die microcontroller chip, printed interconnects are applied over fillets that replace the more standard wire bonds within an IC package.

(in blue, just below and to the left of center in the circuit layout), 2) an LCC-packaged version of a three-axis accelerometer (in blue to the right of center in the circuit layout), 3) LCC versions of both a 5-volt and a 3.3-volt power regulator, and 4) a variety of cavity-mounted resistors, capacitors, and LEDs; a resistor and LED are highlighted by the red box in the right image). [Figure 9\(b\)](#) shows the corresponding, fully fabricated PHE circuit, printed on a flat, 3D printed substrate. PHE fabrication methods not only enable printing onto flat surfaces, but also allow for the fabrication of circuits onto truly 3D surfaces. The circuit design in [figure 10](#) illustrates an earlier version of the PHE circuit projected onto the surface of a 4-inch hemisphere. In the same way that the PHE circuit shown in [figure 9](#) was fabricated, this hemispherical

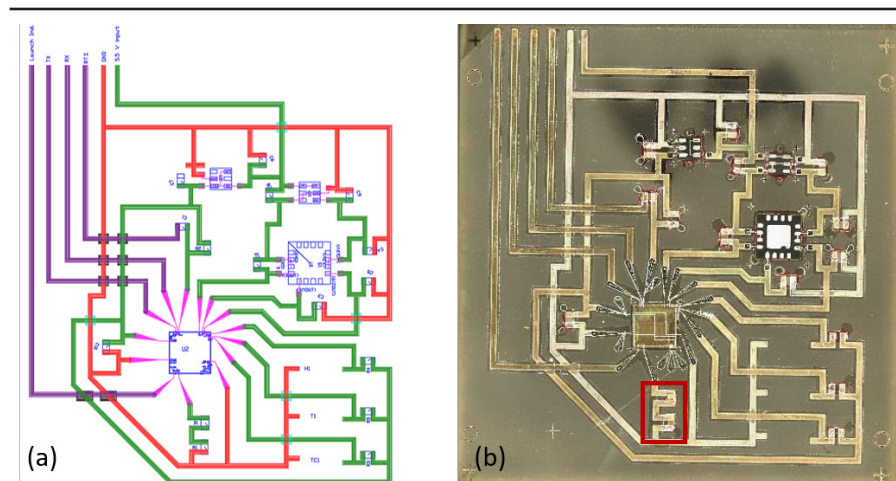


FIGURE 9. The design of a PHE version of an Arduino-type electronic circuit (left) is pictured alongside the fully fabricated AJ printed PHE circuit-level demonstrator (right).

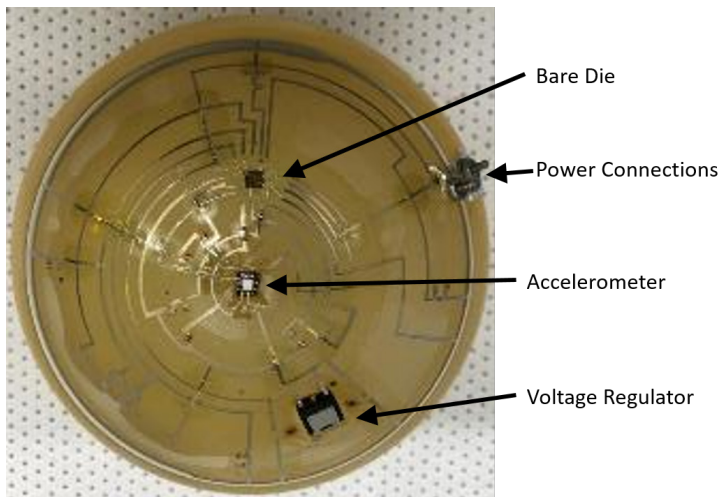


FIGURE 10. Design of a 4-inch hemisphere version of the PHE circuit shown in figure 9 along with the AJ printed, fully fabricated circuit.

circuit was fabricated onto a similarly 3D-printed surface. The only difference in fabrication was that, for the hemispherical circuit shown in figure 10, a five-axis AJ printer was used, while for the flat circuit in figure 9, a three-axis AJ printer was used.

Next steps

As with any new, next-generation technology, proving out reliability and real-world application can be a challenge. This is definitely the case with PHE printing methods used to fabricate 3D, additively fabricated electronic circuits. With this in mind, we are in the process of fabricating some 220 component-level test coupons relevant to the PHE circuit-level demonstrators shown in figure 9 that will go through full reliability testing. Additionally, we are partnering with a number of other government groups, defense industrial base companies, and NextFlex the Manufacturing Innovation Institute (MII) for flexible hybrid electronics (FHE), in order to advance the additive manufacturing ecosystem and prove out the capabilities of this

technology. For example, we are collaborating with NASA [Goddard Space Flight Center (GSFC), Marshall Space Flight Center (MSFC), and their Sounding Rocket Operation Center (NSROC)] to fabricate the PHE circuit onto the inside surface of a door panel for

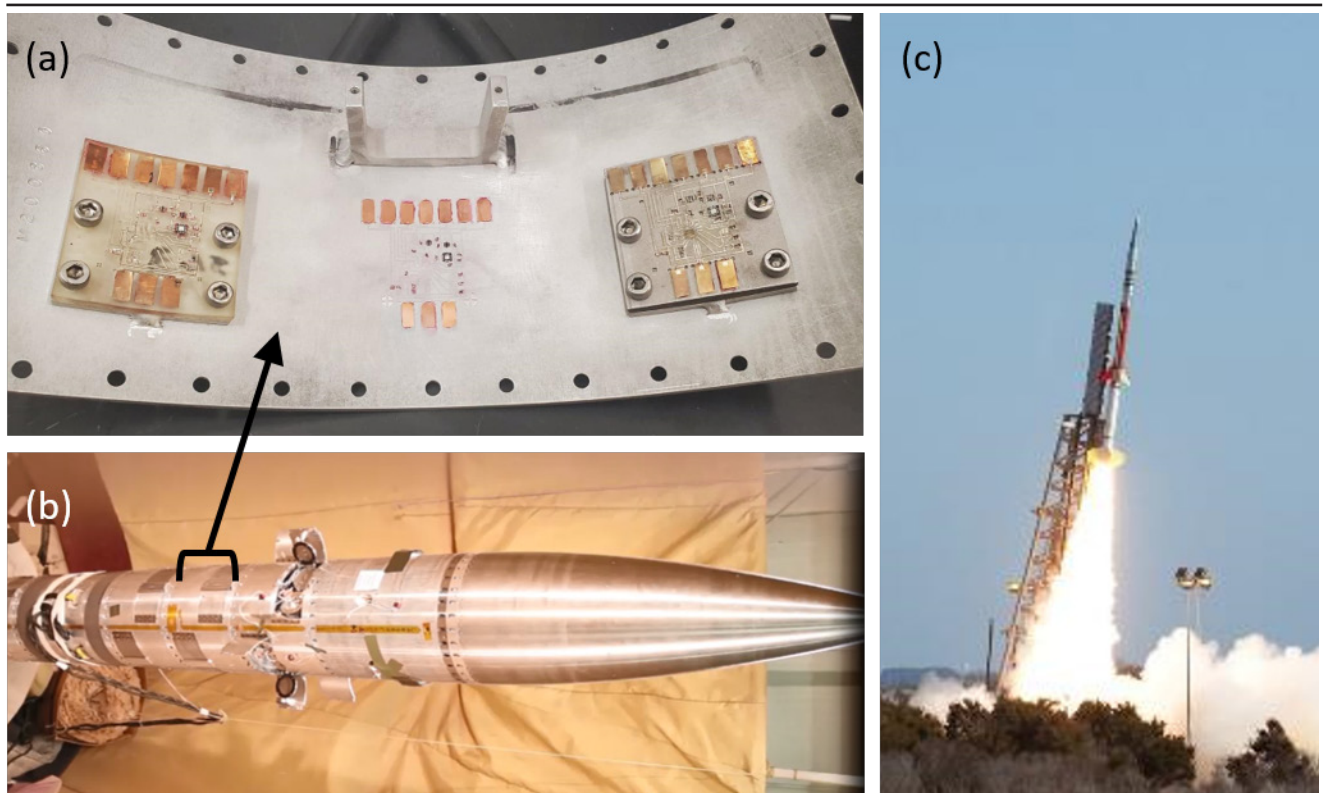



FIGURE 11. (a) The layout of a curved version of the PHE circuit will be fabricated onto the inside surface of a sounding rocket door panel; (b) the sounding rocket is pictured during testing and (c) launching.

a sounding rocket launch in late 2022. A conceptual mock-up is shown in [figure 11](#) with the actual rocket door panel. Included in the figure is a photo of a rocket in test and at launch.

Conclusion

Additive manufacturing holds great promise as a next-generation method for the fabrication of electronic circuits. For one thing, a stand-alone PCB can now be replaced by printing multilayer circuitization onto non-flat surfaces. This allows for the rapid prototyping of circuits that can take on completely different form factors than has been possible in the past. Additionally, printed and hybrid versions of electronic components are typically thinner and lighter weight as compared to their surface-mounted counterparts. This also eliminates the need for soldering, thus further reducing not only weight but also high thermal stress, processing steps, and the number of different materials involved in the overall fabrication process. All in all, the maturity of additive manufacturing methods applied to the fabrication of electronic circuits has the potential to usher in a new era of electronics integration where the circuitry will become inseparable from the mechanical, geometrical aspect of the physical gadget that it controls. 

References

- [1] Huang YA, Wu H, Xiao L, Duan Y, Zhu H, Bian J, Ye D, Yin Z. "Assembly and applications of 3D conformal electronics on curvilinear surfaces." *Materials Horizons*. 2019;6(4):642–683. Available at: <https://doi.org/10.1039/C8MH01450G>.
- [2] Wu H, Tian Y, Luo H, Zhu H, Duan Y, Huang Y. "Fabrication techniques for curved electronics on arbitrary surfaces." *Advanced Materials Technologies*. 2020;5(8):2000093. Available at: <https://doi.org/10.1002/admt.202000093>.
- [3] MacDonald E, Wicker R. "Multiprocess 3D printing for increasing component functionality." *Science*. 2016;353(6307). Available at: <https://doi.org/10.1126/science.aaf2093>.
- [4] Gao W, Zhang Y, Ramanujan D, Ramani K, Chen Y, Williams CB, Wang CC, Shin YC, Zhang S, Zavattieri PD. "The status, challenges, and future of additive manufacturing in engineering." *Computer-Aided Design*. 2015;69:65–89. Available at: <https://doi.org/10.1016/j.cad.2015.04.001>.
- [5] Ligon SC, Liska R, Stampfl J, Gurr M, Mulhaupt R. "Polymers for 3D printing and customized additive manufacturing." *Chemical Review*. 2017;117(15):10212–10290. Available at: <https://doi.org/10.1021/acs.chemrev.7b00074>.
- [6] Espalin D, Muse DW, MacDonald E, Wicker RB. "3D Printing multifunctionality: structures with electronics." *The International Journal of Advanced Manufacturing Technology*. 2014;72:963–978. Available at: <https://doi.org/10.1007/s00170-014-5717-7>.
- [7] Li J, Wasley T, Nguyen T, Ta V, Shephard JD, Stringer J, Smith P, Esenturk E, Connaughton C, Kay R. "Hybrid additive manufacturing of 3D electronic systems." *Journal of Micromechanics Microengineering*. 2016;26(10):105005. Available at: <https://doi.org/10.1088/0960-1317/26/10/105005>.
- [8] Hou Z, Lu H, Li Y, Yang L, Gao Y. "Direct ink writing of materials for electronics-related applications: A mini review." *Frontiers in Materials*. 2021;8:647229. Available at: <https://doi.org/10.3389/fmats.2021.647229>.
- [9] Kwon KS, Rahman MK, Phung TH, Hoath SD, Jeong S, Kim JS. "Review of digital printing technologies for electronic materials." *Flexible and Printed Electronics*. 2020;5(4):043003. Available at: <https://doi.org/10.1088/2058-8585/abc8ca>.
- [10] Wilkinson N, Smith M, Kay R, Harris R. "A review of aerosol jet printing—a non-traditional hybrid process for micro-manufacturing." *The International Journal of Advanced Manufacturing Technology*. 2019;105(11):4599–4619. Available at: <https://doi.org/10.1007/s00170-019-03438-2>.
- [11] Chen G, Gu Y, Tsang H, Hines DR, Das S. "The effect of droplet sizes on overspray in aerosol-jet printing." *Advanced Engineering Materials*. 2018;20(8):1701084. Available at: <https://doi.org/10.1002/adem.201701084>.
- [12] Yoo D, Mahoney CM, Deneault JR, Grabowski C, Austin D, Berrigan JD, Glavin N, Buskohl PR. "Mapping drift in morphology and electrical performance in aerosol jet printing." *Progress in Additive Manufacturing*. 2021;6:257–268. Available at: <https://doi.org/10.1007/s40964-021-00165-7>.
- [13] Tafoya RR, Secor EB. "Understanding and mitigating process drift in aerosol jet printing." *Flex. Print. Electron*. 2020;5(1):015009. Available at: <https://doi.org/10.1088/2058-8585/ab6e74>.
- [14] Gu Y, Gutierrez D, Das S, Hines D. "Ink wells for on-demand deposition rate measurement in aerosol-jet based 3D printing." *Journal of Micromechanics Microengineering*. 2017;27(9):097001. doi: 10.1088/1361-6439/aa817f.

[15] Hines DR, Gu Y, Martin AA, Li P, Fleischer J, Clough-Paez A, Stackhouse G, Dasgupta A, Das S. "Considerations of aerosol-jet printing for the fabrication of printed hybrid electronic circuits." *Additive Manufacturing*. 2021;47:102325. Available at: <https://doi.org/10.1016/j.addma.2021.102325>.

[16] Tafoya RR, Cook AW, Kaehr B, Downing JR, Hersam MC, Secor EB. "Real-time optical process monitoring for structure and property control of aerosol jet printed functional materials." *Advanced Materials Technologies*. 2020;5(12):2000781. Available at: <https://doi.org/10.1002/admt.202000781>.

[17] Yi C, Fedderwitz R, Park D, Ding C, Lu G-Q, Fleischer J, Li P, Kofinas P, Das S, Hines DR. "Fully printed resonance-free broadband conical inductors using engineered magnetic inks." *Additive Manufacturing*. 2021;44:102034. Available at: <https://doi.org/10.1016/j.addma.2021.102034>.

[18] Khan Y, Garg M, Gui Q, Schadt M, Gaikwad A, Han D, Yamamoto NA, Hart P, Welte R, Wilson W. "Flexible hybrid electronics: Direct interfacing of soft and hard electronics for wearable health monitoring." *Advanced Functional Materials*. 2016;26(47):8764–8775. Available at: <https://doi.org/10.1002/adfm.201603763>.

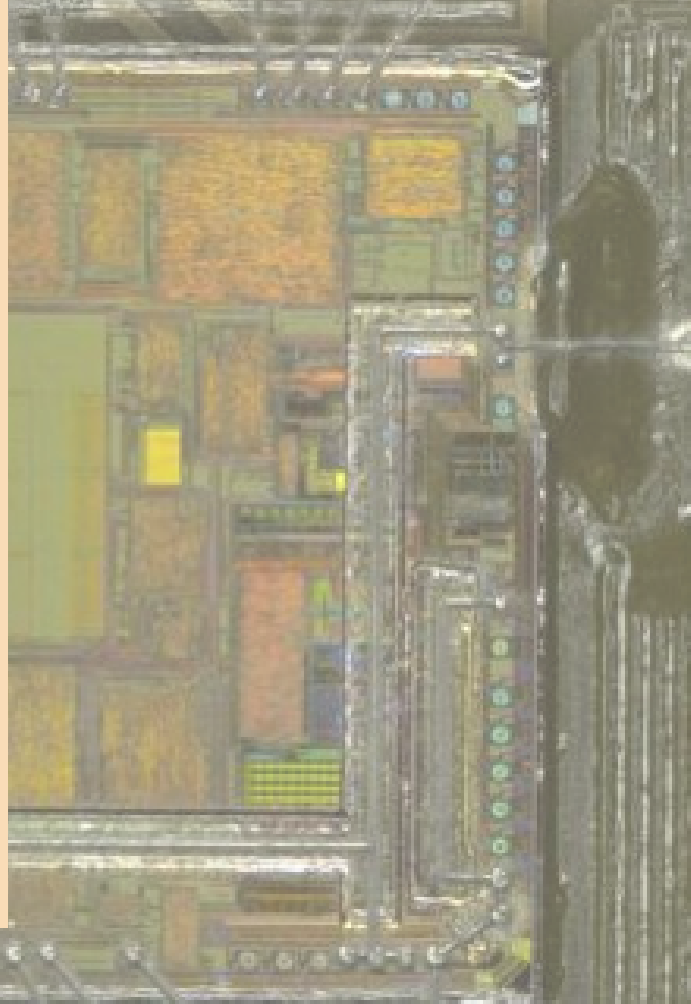
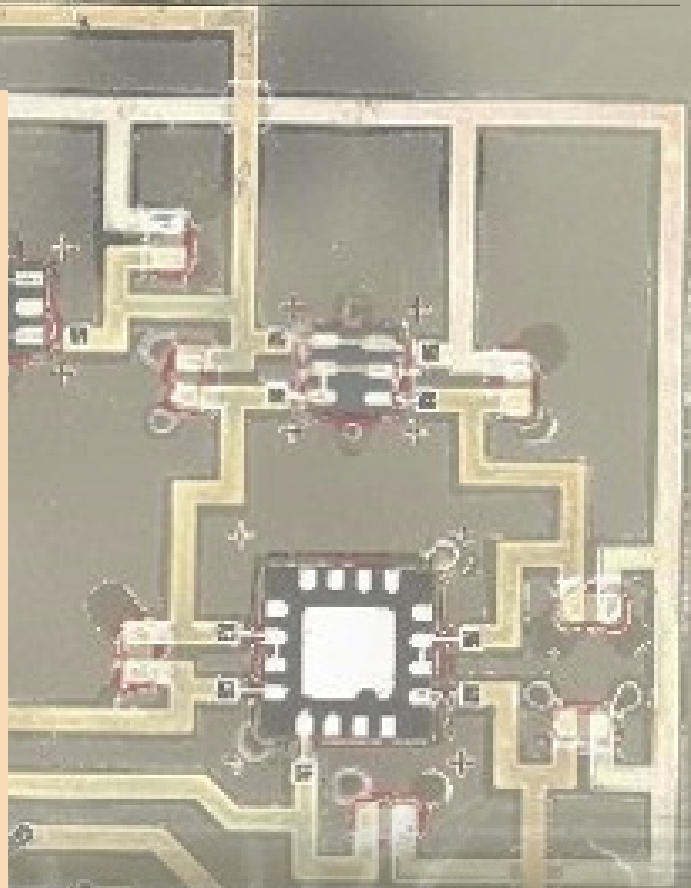
[19] Khan Y, Thielens A, Muin S, Ting J, Baumbauer C, Arias AC. "A new frontier of printed electronics: Flexible hybrid electronics." *Advanced Materials*. 2020;32(15):1905279. Available at: <https://doi.org/10.1002/adma.201905279>.

[20] Hoerber J, Glasschroeder J, Pfeffer M, Schilp J, Zaeh M, Franke J. "Approaches for additive manufacturing of 3D electronic applications." *Procedia CIRP*. 2014;17:806–811. Available at: <https://doi.org/10.1016/j.procir.2014.01.090>.

[21] Gupta AA, Bolduc A, Cloutier SG, Izquierdo R. "Aerosol jet printing for printed electronics rapid prototyping." In: *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*; 2016 May 22–25; Montreal, QC, Canada: pp. 866–869. Available at: <https://doi.org/10.1109/ISCAS.2016.7527378>.

[22] Clough-Paez A, Yi C, Park D, Ketchum D, Hines DR. "Rapid prototyping of 3D printed, high aspect ratio, low noise amplifier for active hand-held sensor devices." Submitted to *IEEE Transactions on Components, Packaging and Manufacturing Technology* for publication (2022).

[23] Gu Y, Hines DR, Yun V, Antoniak M, Das S. "Aerosol-Jet printed fillets for well-formed electrical connections between different leveled surfaces." *Advanced Materials Technologies*. 2017;2(11):1700178. Available at: <https://doi.org/10.1002/admt.201700178>.





[Photo credit: iStock.com/EasternLighcraft]

Selected Publications by NSA Researchers, 2021–2022

The NSA Research Directorate is the largest in-house research organization in the Intelligence Community, with experts in fields such as mathematics, computer science, engineering, cybersecurity, physics, neuroscience, and linguistics. Our researchers are active in their fields within and outside of NSA. The following bibliography lists selected publications from January 1, 2021 to December 31, 2022 for which one or more of the authors are NSA researchers.

1. Baron JD, Darling RWR, Davis JL, Pettit R. "Partitioned K-nearest neighbor local depth for scalable comparison-based learning." 2021. Cornell University Library. Available at: <https://doi.org/10.48550/arXiv.2108.08864>.
2. Baron JD; Darling RWR. "Empirical complexity of comparator-based nearest neighbor descent." 2022. Cornell University Library. Available at: <https://doi.org/10.48550/arXiv.2202.00517>.
3. Bhandary P, Ziegler E, Nicholas C. "Searching for selfie in TLS 1.3 with the cryptographic protocol shapes analyzer." 2021. In: Dougherty D, Meseguer J, Mödersheim SA, Rowe P (eds), *Protocols, Strands, and Logic. Lecture Notes in Computer Science, vol 13066*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-91631-2_3.
4. Bilinski M, diVita J, Ferguson-Walter K, Fugate S, Gabrys R, Mauger J, Souza B. (2021). "No time to lie: Bounds on the learning rate of a defender for inferring attacker target preferences." 2021. In: Bošanský B, Gonzalez C, Rass S, Sinha A (eds), *Decision and Game Theory for Security. GameSec 2021. Lecture Notes in Computer Science, vol 13061*. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-90370-1_8.

5. Bluher A. "A new identity of Dickson polynomials." *Finite Fields and Their Applications*. 2022;80(102012). Available at: <https://doi.org/10.1016/j.ffa.2022.102012>.
6. Bluher A. "Explicit Artin maps into PGL2." *Expositiones Mathematicae*. 2022;40(1):45–93. Available at: <https://doi.org/10.1016/j.exmath.2021.07.003>.
7. Bluher A. "New Wilson-like theorems arising from Dickson polynomials." *Finite Fields and Their Application*. 2021;72(101819). Available at: <https://doi.org/10.1016/j.ffa.2021.101819>.
8. Bluher A. "Permutation properties of Dickson and Chebyshev polynomials with connections to number theory." *Finite Fields and Their Application*. 2021;76(101899). Available at: <https://doi.org/10.1016/j.ffa.2021.101899>.
9. Brewster RA, Goldhar J, Morris M, Baumgartner G, Chembo Y. "Estimation of the CHSH parameter using HOM interference." *IEEE Transactions on Quantum Engineering*. 2022;3(4100310):1–10. Available at: <https://doi.org/10.1109/TQE.2022.3152170>.
10. Burkhardt P. "Graph connectivity in log steps using label propagation." *Parallel Processing Letters*. 2021;31(4):1–23. Available at: <https://doi.org/10.1142/S0129626421500213>.
11. Burkhardt P. "Optimal algebraic breadth-first search for sparse graphs." *ACM Transactions on Knowledge Discovery from Data*. 2021;15(5):1–19. Available at: <https://doi.org/10.1145/3446216>.
12. Comar B. "Decoding and equalization of a joint Alamouti-WPM system." In: *IEEE International Symposium on Networks, Computers, and Communications (ISNCC), 2021*. Available at: <https://doi.org/10.1109/ISNCC52172.2021.9615704>.
13. Comar B. "Detection of sinusoids with frequency drift in white Gaussian noise." *IEEE International Conference on Information and Communication Technology (ICoICT), 2021*. Available at: <https://doi.org/10.1109/ICoICT52021.2021.9527437>.
14. Comar B. "Implementation of a QPSK symbol synchronizer in Xilinx system generator," In: *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2021*, pp. 396–401. Available at: <https://doi.org/10.1109/ISVLSI51109.2021.00078>.
15. Comar B. "Method of combining cryptography and LDPC coding for enhanced privacy." *IGI Global: International Journal of Interdisciplinary Telecommunications & Networking (IJITN)*. 2021;13(4). Available at: <https://doi.org/10.4018/IJITN.2021100105>.
16. Comar B. "Tone detection system design for targets with frequency drift." *IEEE International Conference on Information and Communication Technology (ICoICT), 2021*. Available at: <https://doi.org/10.1109/ICoICT52021.2021.9527412>.
17. Comar B, Frazier S. "Design of a joint alamouti-MIMO and wavelet packet modulation system." In: *IEEE International Symposium on Networks, Computers, and Communications (ISNCC), 2021*. Available at: <https://doi.org/10.1109/ISNCC52172.2021.9615650>.
18. Darling RWR. "Hidden ancestor graphs with assortative vertex attributes" 2021. Cornell University Library. Available at: <https://doi.org/10.48550/arXiv.2102.09581>.
19. Darling RWR. "Hidden ancestor graphs with assortative vertex attributes." *Scaling limits: From statistical mechanics to manifolds. A workshop in honor of James Norris' 60th birthday*. 2022. Available at: [dx.doi.org/10.13140/RG.2.2.14978.56006](https://doi.org/10.13140/RG.2.2.14978.56006).
20. Darling RWR, Emanuello JA, Purvine E, Ridley A (eds). *Proceedings of TDA: Applications of Topological Data Analysis to Data Science, Artificial Intelligence, and Machine Learning Workshop at SDM 2022*. Cornell University Library. Available at: <https://doi.org/10.48550/arXiv.2204.01142>.
21. Edwards CA, Goyal A, Rusheen AE, Kouzani AZ, Lee KH. "DeepNavNet: Automated landmark localization for neuronavigation." *Frontiers in Neuroscience*. 2021;15. Available at: <https://doi.org/10.3389/fnins.2021.670287>.

22. Ferguson-Walter KJ, Major MM, Johnson CK, Muhleman DH. "Examining the efficacy of decoy-based and psychological cyber deception." In: *Proceedings of the 30th USENIX Security Symposium*; 2021 Aug 11–13. Available at: <https://usenix.org/system/files/sec21-ferguson-walter.pdf>.
23. Gerstner CR, Farid H. "Detecting real-time deep-fake videos using active illumination." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 53–60. Available at: <https://doi.org/10.1109/CVPRW56347.2022.00015>.
24. Golczynski A, Emanuello J. "End-to-end anomaly detection for identifying malicious cyber behavior through NLP-based log embeddings." 2021. Paper presented at the *International Joint Conference on Artificial Intelligence (IJCAI) First International Conference on Adaptive Cyber Defense*. Available at: <https://doi.org/10.48550/arXiv.2108.12276>.
25. Johnson C, Ferguson-Walter K, Gutzwiller R, Scott D, Cook N. "Investigating cyber attacker team cognition." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2022;66(1):105–109. Available at: <https://doi.org/10.1177/1071181322661132>.
26. Khan A, Fleming E, Schofield N, Bishop M, Andrews N. 2021. "A deep metric learning approach to account linking." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5275–5287. Available at: <http://dx.doi.org/10.18653/v1/2021.naacl-main.415>.
27. Laird B, Tran T. "Data-informed CRLB derivations for indoor emitter localization." In: *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021, pp 509–516. Available at: <https://doi.org/10.1109/IEEECONF53345.2021.9723413>.
28. Laird B, Tran T. "Quasi-norm kernel-based emitter localization." In: *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 2021, pp 534–538. Available at: <https://doi.org/10.1109/IEEECONF53345.2021.9723416>.
29. Major MM, Souza BJ, DiVita J, Ferguson-Walter KJ. "Informing autonomous deception systems with cyber expert performance data." 2021. Paper presented at the *International Joint Conference on Artificial Intelligence (IJCAI) First International Conference on Adaptive Cyber Defense*. Available at: <https://doi.org/10.48550/arXiv.2109.00066>.
30. Medak A. "Hardening the hardware supply chain: Standardized artifacts enable automated accountability." *The Next Wave*. 2022;23(2):40–49. ISSN 2640-1789. Available at: <https://www.nsa.gov/thenextwave>.
31. Molina-Markham A, Minter C, Powell B, Ridley A. "Network environment design for autonomous cyberdefense." 2021. Cornell University Library. Available at: <https://doi.org/10.48550/arXiv.2103.07583>.
32. Molina-Markham A, Winder RK, Ridley A. "Network defense is not a game." 2021. Paper presented at *2021 SIAM International Conference on Data Mining Workshop on AI/ML for Cybersecurity: Challenges, Solutions and Novel Ideas*. Available at: <https://doi.org/10.48550/arXiv.2104.10262>.
33. Nguyen AT, Raff E, Nicholas C, Holt J. "Leveraging uncertainty for improved static malware detection under extreme false positive constraints." 2021. Paper presented at the *International Joint Conference on Artificial Intelligence (IJCAI) First International Conference on Adaptive Cyber Defense*. Available at: <https://doi.org/10.48550/arXiv.2108.04081>.
34. Pellicone A, Ketelhut D, Shokeen E, Weintrop D, Cukier M, Plane J. "Designing a game to promote equity in cybersecurity." *Proceedings of the 16th European Conference on Games Based Learning*. 2022;16(1). Available at: <https://doi.org/10.34190/ecgbl.16.1.825>.
35. Pellicone A, Shokeen E, Moon P, Weintrop D, Ketelhut D, Cukier M, Plane J. "It just felt more like a pyramid—Narrative and concept in game-based learning puzzles." *2022 International Conference of Meaningful Play*. Abstract available at: <https://meaningfulplay.msu.edu/proceedings2022/>.

36. Rivera-Soto RA, Miano OE, Ordonez J, Chen BY, Khan A, Bishop M, Andrews N. "Learning universal authorship representations." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 913–919, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. Available at: <http://dx.doi.org/10.18653/v1/2021.emnlp-main.70>.
37. Rolinger TB, Krieger CD, Sussman A. "Optimizing memory-compute colocation for irregular applications on a migratory thread architecture." In: *2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2021, pp. 58–67. Available at: <https://doi.org/10.1109/IPDPS49936.2021.00015>.
38. Rolinger TB, Krieger CD, Sussman A. "Runtime optimizations for irregular applications in Chapel." In: *8th Annual Chapel Implementers and Users Workshop (CHI UW), 2021*. Available at: <https://chapel-lang.org/CHI UW/2021/RolingerSlides.pdf>.
39. Shokeen E, Pellicone A, Weintrop D, Moon P, Cukier M, Ketelhut D, Plane J. "The game was designed to learn to think—Player perception of learning in an educational game." In: *Proceedings of the 16th International Conference of the Learning Sciences (ICLS) 2022*. Available at: <https://2022.isls.org/proceedings/>.
40. Walter EC, Ferguson-Walter KJ, Ridley AD. "Incorporating deception into CyberBattleSim for autonomous defense." 2021. Paper presented at the *International Joint Conference on Artificial Intelligence (IJCAI) First International Conference on Adaptive Cyber Defense*. Available at: <https://doi.org/10.48550/arXiv.2108.13980>.
41. Wong V, Emanuella J. "Robustness of ML-enhanced IDS to stealthy adversaries." 2021. Paper presented at *2021 SIAM International Conference on Data Mining Workshop on AI/ML for Cybersecurity: Challenges, Solutions and Novel Ideas*. Available at: <https://doi.org/10.48550/arXiv.2104.10742>.

Careers **with** **NATIONAL IMPACT**

Work where continual learning, work-life balance and contributing to the greater good are top priorities.

NSA keeps the nation safe through our signals intelligence and cybersecurity missions. If you have a heart for service, there's a place for you here.

**Apply now at
IntelligenceCareers.gov/NSA**



**NSA
Careers**



U.S. citizenship required. NSA is an Equal Opportunity Employer.

**Business & Contracting
Cybersecurity
Foreign Language
Intelligence
STEM
Professional Support**





NATIONAL SECURITY AGENCY



CENTRAL SECURITY SERVICE

Defending Our Nation. Securing The Future.