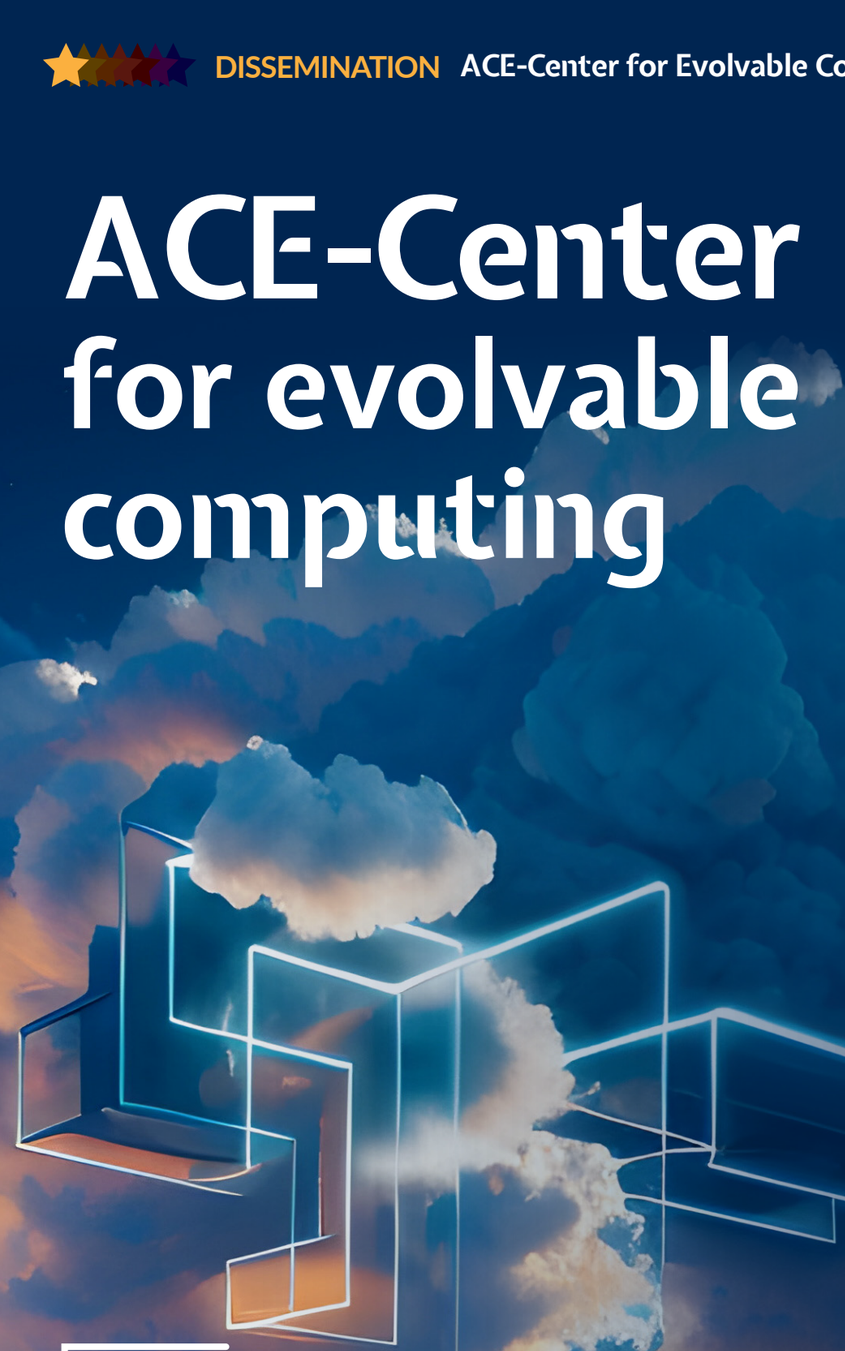
ACE-Center for evolvable computing



How new technologies can make our computers faster and use less energy when working together.

Adobe Stock © *GridsAndTiles*

The computing demands of our society will continue to grow exponentially throughout this decade (Figure 1). Web search, content distillation and recommendation systems, and AI-based content generation services will continue to expand. Increasing amounts of computation will be spent on supporting extended reality with real-time 3D immersion. The rise of autonomous vehicles, drones and robotic assistants, together with the internet of things (IoT) (Wikipedia Contributors, 2023b), will further drive computing needs.

Based on these trends, it is clear that distributed computing in the next decade will be defined by the need to process vast swaths of data for insights in a timely manner. Curtailing energy consumption through data movement minimisation and efficient computation will be the overriding constraint. The compute infrastructure will be a seamless hierarchy of compute centres from edge centres to geo-distributed mega-datacentres. For energy efficiency, each compute centre will contain a large number of heterogeneous hardware accelerators, and tasks of unprecedentedly small granularity will transparently ship computation to where data is. The computational environment will be highly dynamic, with the constant introduction of new classes of compute accelerators for barely-emerging workloads and of new applications or protocols that could benefit from yet-to-be-conceived accelerators.

Given this landscape, the goal of the ACE Center is to devise novel technologies for scalable computing that will substantially improve the performance and energy efficiency of distributed computing in the next decade. To do so, we are exploring a set of new processing, storage, communication and security technologies that operate in a tightly-coupled manner.

To attain our goal, we argue for an *evolvable computing paradigm*. The idea is that the new accelerator hardware, memory structures, communication stacks and security mechanisms need to be designed for extensibility and composability. Specifically, components such as hardware accelerators, memories and interconnects should have standard and composable interfaces, so they can be easily assembled into systems of different form factors, survive upgrades of their external environments, and be easily replaced by (and co-exist with) their next-generation designs. These properties have served us well in the era of general-purpose processors; as we now move to an accelerator-intensive era, we need to retain them.

At the same time, applications should be built as collections of functions that abstract the details of the accelerator they run on and the communication mechanisms they use. With this approach, we will attain not only evolvability but faster time to deployment as well. Such a transition is already taking place with new frameworks like microservices.

Figure 2 depicts ACE's vision of the planet-scale distributed computing infrastructure of the next decade.

Figure 1: Application domains that are driving computation.

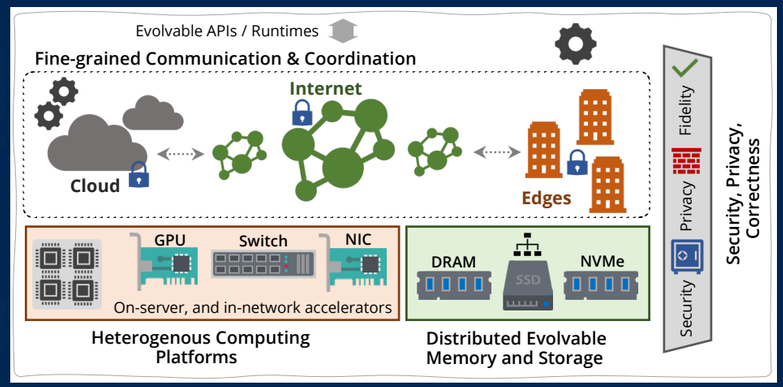


Figure 2: Components of the 2030 planet-scale distributed computing infrastructure envisioned by the ACE Center.

A computing backbone based on accelerators

An integral part of the computing infrastructure of the next decade will be a myriad of different types of accelerators. Accelerators can attain orders of magnitude increases in performance or energy efficiency by eliminating (or repurposing) some of the functionality in different layers of the computing stack—intuitively, by compressing the stack. Accelerators will rapidly evolve with applications and, in addition, at any point in time, co-exist with earlier or later generations. To reach this vision, we are developing a new methodology to easily generate, deploy and reconfigure evolvable accelerators.

To attain ACE's goals, a few types of accelerators will not suffice. Instead, we will need to use multiple accelerators to speed up even a single application. To address this challenge, we are developing technologies to design *ensembles* of accelerators. Some ensembles will be centralised into a chiplet-based package

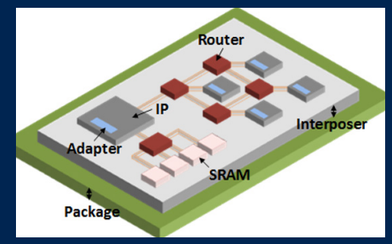


Figure 3: Ensemble of accelerators in a package.

(Figure 3) or a few nearby packages. In most cases, however, an ensemble will be distributed across a data centre, and the accelerators will be connected with high-speed electrical or optical links. Then, applications will select and use the desired set of accelerators from the accelerator ensemble.

No amount of hardware smarts alone will be able to attain this vision. Hence, we are also developing robust runtime and compilation methods that enable quick reconfiguration of the accelerator ensemble and mapping and scheduling of applications to the ensemble. Importantly, since these accelerator ensembles will be spatially and temporally shared by multiple tenants, we are developing new security and privacy mechanisms for multitenancy.

General-purpose cores will have a major role to play, as they are inexpensive and handle all kinds of workloads. However, some of their mechanisms will need to be revamped to scale to the new environments. We believe that general-purpose cores will come in a variety of specialised configurations that will allow them to adapt to different environments, thus substantially increasing their performance and energy efficiency (Stojkovic *et al.*, 2023).

Heterogeneous and intelligent memory and storage

Extrapolation of current trends suggests that memory and storage systems will be complex and highly heterogeneous. Currently, solid state drives (SSDs) are bringing storage closer to memory, non-volatile memory (NVM) is blurring the

line between storage and memory, and new memory interconnects like stacked DRAM (Wikipedia Contributors, 2023a) are driving up memory performance. New disruptive technologies will continue to appear.

As workloads like AI training relentlessly increase their data needs, Compute Express Link (CXL)-like mechanisms (Compute Express Link, 2022) will help expand the abstraction of local memory beyond a node to possibly an entire rack. The result will be a formidable memory wall that will demand novel mechanisms for processors to tolerate latency and for data-coherence performance to degrade gracefully with distance.

To utilise these heterogeneous memory and storage resources efficiently, we will need abstractions to designate the characteristics of memory and storage assets so that applications can select the type of asset they need. Moreover, we are working on theory-grounded scalable algorithms that apportion datacenter-scale assets fairly among thousands of competing applications and potentially billions of allocation requests.

With the advent of intelligent memory and storage (IMS) (Liu *et al.*, 2021), distributed computers will have ubiquitous IMS blocks in many parts of the memory hierarchy, including possibly every level of the caches, DRAM and NVM modules in main memory, network switches and SSDs. We are developing hardware and software techniques to harness these distributed IMS blocks so they operate in a coordinated manner.

Fine-grained communication and coordination

Computing and storage systems will communicate via high-performance, energy-efficient networks. Geo-distributed data centres with millions of accelerators will emerge (Figure 4) and support a wide variety of heterogeneous functionality. They will have reconfigurable network topologies that will leverage accelerators for protocol and infrastructure tasks.

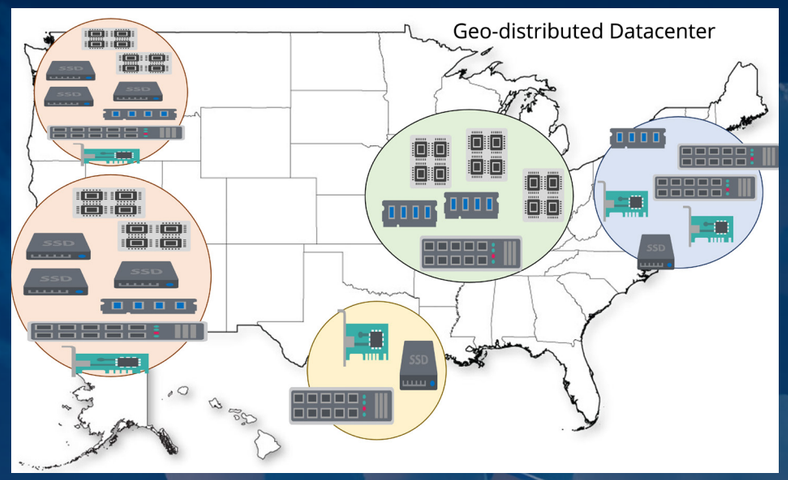


Figure 4: Geo-distributed data centre.

Efficiency in these data centres will not be attained with advanced hardware only; the communication software stack needs to be redesigned as well. Current stacks were designed assuming a general-purpose homogeneous environment and support all possible functionality. Instead, we are developing an evolvable communication software stack that specialises in the accelerators available—substantially reducing the *data centre tax*.

Even with the most advanced accelerators and lean communication stacks, we will not attain the performance gains we seek if these accelerators are often sitting idle because the scheduler fails to assign them work. Moreover, we will not attain the energy reductions we seek if these accelerators often operate on remote data, since the energy consumed by data movement will remain dominant. To make a significant difference, we envision changes in the way computation is packaged, migrated and scheduled. We are designing a new runtime that bundles computation in small buckets that can be quickly migrated and shipped to where the data lives (Ruan *et al.*, 2023).

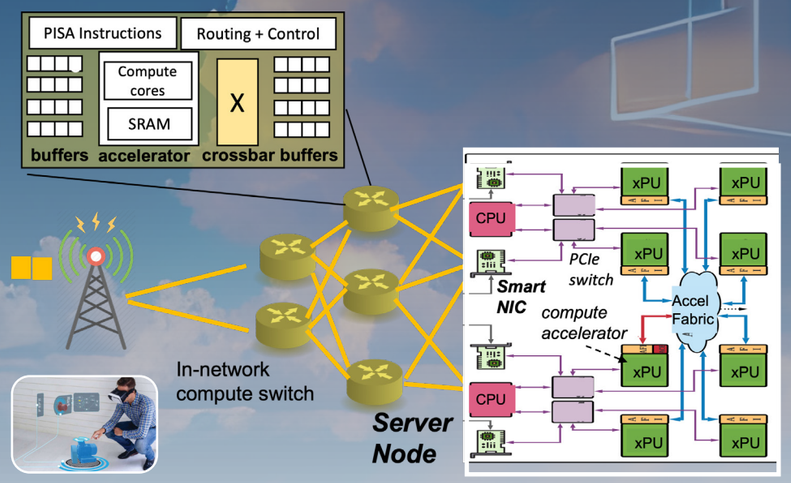


Figure 5: Accelerators in network switches and smart network interface cards (SmartNICs).

Accelerators in network switches and network interface cards (NICs) will use their unique vantage points to perform a variety of operations more efficiently than processors (Figure 5). Such computation offloading will further improve efficiency.

Secure, private and correct accelerators

Distributed systems will increasingly process sensitive data. Currently, the large majority of security frameworks tie security properties to users, applications, or hardware/software systems. We believe that such approaches are intrinsically fragile. Hence, we are developing new *data-centric* security paradigms that tie the security properties to data. As data moves across planet-scale infrastructure, the data's security properties hold.

The move toward an ecosystem rich in accelerators does call for a robust methodology to build accelerators with security and correctness mechanisms from the ground up. Consequently, we envision extending trusted execution environment (TEE) designs to accelerators. For highest efficiency, these TEEs will be customised to the particular accelerator, be evolvable to adapt to upgrades and changes in the environment and, importantly, be generated automatically by a compiler and a common framework.

We are adapting key methodologies and tools currently used to verify the security properties of processors to apply to accelerators. Examples of these methodologies are hardware-level information flow control (IFC), register transfer level (RTL) analyses to discover security vulnerabilities, and leakage-detecting techniques inspired by hardware verification. Finally, as accelerators aim for short design-to-deployment timelines, we are developing new design-for-verification (DFV) principles for future evolvable accelerators.

Final remarks

The mission of the ACE Center is to make this vision a reality. We hope to put the entire computing ecosystem on a new trajectory of system evolution with much higher performance and energy efficiency improvement.

References

Compute Express Link (2022) *Compute Express Link: The Breakthrough CPU-to-Device Interconnect*'. Available at: https://www.computeexpresslink.org.

Liu, K., Zhang, X., So, J., Lee, J.-G., Kang, S.-H., Lee, S., Han, S., Cho, Y., Kim, J.H., Kwon, Y., Kim, K., Jung, J., Yun, I., Park, S.J., Park, H., Song, J., Cho, J., Sohn, K., Kim, N.S. and Lee, H.-H.S. (2021) 'Near-Memory Processing in Action: Accelerating Personalized Recommendation with AxDIMM.' *IEEE Micro*, 42(1) pp. 116–127. doi: 10.1109/MM.2021.3097700.

Ruan, Z., Park, S.J., Aguilera, M.K., Belay, A., Schwarzkopf, M. (2023) 'Nu: Achieving microsecond-scale resource fungibility with logical processes' in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, Boston, MA, pp. 1409–1427.

Stojkovic, J., Liu, C., Shahbaz, M., Torrellas, J. (2023) 'µManycore: A Cloud-Native CPU for Tail at Scale.' *International Symposium on Computer Architecture (ISCA), Orlando, FL.*

Wikipedia Contributors (2023a) *High Bandwidth Memory*. Available at: https://en.wikipedia.org/wiki/High_Bandwidth_Memory.

Wikipedia Contributors (2023b) 'Internet of Things'. Available at: https://en.wikipedia.org/wiki/Internet_of_things.

ACE CENTER FOR EVOLVABLE COMPUTING

PROJECT SUMMARY

The aim of the ACE Center is to devise novel technologies that will substantially improve the performance and energy efficiency of distributed computing in the next decade. ACE innovates processing, storage, communication, and security technologies that address the seismic shifts identified in the Semiconductor Research Corporation (SRC) Decadal Plan for Semiconductors.

PROJECT PARTNERS

The ACE team: Josep Torrellas (Director, Univ. Illinois), Minlan Yu (Assistant Director, Harvard), Tarek Abdelzaher (Univ. Illinois), Mohammad Alian (Univ. Kansas), Adam Belay (MIT), Manya Ghobadi (MIT), Rajesh Gupta (UCSD), Christos Kozyrakis (Stanford), Tushar Krishna (GaTech), Arvind Krishnamurthy (Univ. Washington), Jose Martinez (Cornell), Charith Mendis (Univ. Illinois), Subhasish Mitra (Stanford), Muhammad Shahbaz (Purdue), Edward Suh (Cornell), Steven Swanson (UCSD), Michael Taylor (Univ. Washington), Radu Teodorescu (Ohio State Univ.), Mohit Tiwari (Univ. Texas), Zhengya Zhang (Univ. Michigan) and Zhiru Zhang (Cornell).

PROJECT LEAD PROFILE

Josep Torrellas is the Saburo Muroga Professor of Computer Science at University of Illinois Urbana-Champaign. His research interests are parallel computer architectures. He has contributed to several experimental multiprocessor designs such as IBM's PERCS Multiprocessor, Intel's Runnemede Extreme-Scale Multiprocessor, Illinois Cedar and Stanford DASH. He is a Fellow of IEEE, ACM and AAAS. He received a PhD from Stanford University.

PROJECT CONTACTS

Josep Torrellas
201 N. Goodwin Avenue, Urbana, IL, 61801, USA
+1-217-979-7820
torrella@illinois.edu
https://acecenter.grainger.illinois.edu
https://twitter.com/ace_computing
https://www.linkedin.com/company/ace-center-for-evolvable-computing

FUNDING

ACE is a JUMP 2.0 Center sponsored by the Semiconductor Research Corporation (SRC) and the Defense Advanced Research Projects Agency (DARPA).