Revolutionizing the Digital World: Unveiling the Future of Computing with Spatial Awareness, Generative AI, and Dynamic Web Integrations!

The next computing paradigm: an introduction

by Tullio Vardanega and Marc Duranton

What will be the future of computing systems (infrastructure, software and hardware)? HiPEAC envisions *the next computing paradigm* (NCP), focusing on a seamless integration of key ingredients from various digital elements like the Web, the Cloud, Cyber-Physical Systems, the Internet of Things, digital twins, the metaverse, and Artificial Intelligence. Envisioning the NCP emphasizes the evolution towards a spatial dimension in computing, a coherent continuum of computing, intertwining the real world with the cyberworld, incorporating Generative AI, and dynamic orchestrations of resources. The aim is to create a seamless, networked cooperative structure where resources are accessed and manipulated with streamlined Web-type protocols, where programs (in fact services) and data flow smoothly onto computing resources that cooperate with each other enhancing context-awareness and efficiency in digital interactions.

Key insights

- Integration of Digital Elements: This dimension combines the Web, the Cloud, Cyber-Physical Systems, the Internet of Things, digital twins, the metaverse, and Artificial Intelligence, for a holistic digital experience.
- **Spatial Computing:** This dimension adds a spatial-awareness trait to digital interactions, considering physical constraints and location-dependent factors as well as time, resulting in (at least) a 4D computing paradigm.
- **Generative AI at the Edge:** This dimension embeds personalized AI engines in edge devices for dynamic, on-the-fly construction of smart ad-hoc orchestrated hoc applications.
- **Dynamic Web Integrations:** This dimension focuses on ephemeral, context-aware aggregations based on web resources, improving accessibility and efficiency.
- **Shift Towards User-Centric Models:** This dimension moves computation closer to users or data sources to balance latency, privacy, and energy requirements.

Key recommendations

Enacting the Next Computing Paradigm vision outlined in this part of the HiPEAC vision requires a number of distinct demanding technology improvements, each of which shall pay great attention to interoperability:

- Developing stacks of 4D-aware implementation technologies capable of spatial and time-aware computing. Doing so will entail merging streamlined evolutive versions of HTTP-based REST-like web protocols with Spatial Web protocols (HSML and HSTP, OpenUSD), and augmenting them so that they can guarantee timely delivery for any granularity of time.
- Augmenting APIs with contract-based interoperable specifications that enable ephemeral (on-the-fly) compositions based on the

pairing of assume/guarantee declarations across required and provided interfaces.

- Allowing computation to move after traveling users or objects, or toward specific data sources, seeking best balance among latency, privacy, data freshness and provenance, and energy requirements. Seeking this objective will require improving WebAssembly-type technology capable of enabling sandboxed hosting and efficient (interpreted) execution of in-transit bundles of computations.
- Developing AI-powered Edge-based trustworthy (robustly loyal) orchestrators that dynamically, opportunistically, and ephemerally assemble remote APIs into ad-hoc private service compositions.

Our modern world

If we wanted to enumerate the key elements of the digital space that surrounds us as individuals, professionals, and members of the social fabric in which we live, we would likely agree on the following list of items:

- The **Web**, as the infrastructure that supports most of our activities over the Internet. The Internet, inaugurated about 40 years ago, was the primary enabler to worldwide connectivity. The Web came some 10 years afterwards, and progressively changed just about everything as far as everyone's experience of networking goes. If we corresponded our "navigating the network" to moving around a (gigantic and virtual) building, the Internet would be its foundation, so much below ground to be invisible, and the infrastructure of the building, which holds all contents together and allow users to move conveniently around them, would be the Web. So the Web infrastructure – its protocols and way of use – is essential to where we stand today.
- The **Cloud**, probably the most impactful byproduct of the Web to date, which renders each digital thing available as a web resource in an as-a-service mode. The concept of Cloud originated from the visionary realization that everything could be exposed and access as a web resource, not only static data, but also computation (apps), and computing resources (CPUs, storage, networking). Remarkably, this vision was put forward in 1969 (!) by one of the founders of the ARPANET that predated the subsequent Internet. Intrinsic to the realization of that vision was the Web infrastructure that we have come to know as "the Cloud", thus becoming the most global and ubiquitous programming-and-execution platform that ever existed. Very much in line with the Web logic of access to the resources exposed in it, the as-a-service model that characterizes the Cloud allows tapping applications without needing to install them locally (This should be no surprise: when you "navigate" to a resource, you consume it locally to your home base, but you know very well that it stays where it was). As part of that innovation, the REST architectural style made HTTP – the principal enabler of the design of modern Web apps – the means to access and manipulate web resources programmatically in a uniform and consistent way. REST is a most natural and productive way to leverage the as-a-service style of the Cloud: with REST, higher-level services can easily be realized by orchestrating others, in a most versatile value-added way. So we can say that the HTTP-based and RESTful view of the Web infrastructure is the foundation to modern digital services and applications.
- The **Internet of Things**, which timidly originated from equipping non-digital items with radio-frequency identification devices that would allow them to be interrogated digitally, if only for tracking purposes. Soon after, that rudimentary concept evolved into requiring such items to become "smart", thus, capable of sensing and actuation, and sometimes even of basic in-place processing, eventually interconnecting them with human-side devices or among themselves. The IoT has become a rich and pervasive mesh of connected digital "things", which allows mission-specific value-added services to be provided, to various types of target groups, up to entire populations, as it would be in a "smart city". So we can say that the digital means (protocols and software infrastructures) to get things smart and connected is a transformative convenience to lots of our daily activities, whether professional, social, or personal.
- **Cyberphysical Systems**, which can be seen as the command-and-control processing part of all sorts of articulations of IoT devices deployed into mission-critical products that help us build "intelligent" industrial and civil infrastructures. CPSs are perhaps not very conspicuous to the general population, but they are found at all places where controlled

automation is needed to guarantee the delivery of critical services in transportation, health, manufacturing, and a growing number of other sectors. As CPSs control physical devices, safety concerns arise, together with security concerns, which are common to all other critical ICT infrastructures. The central tenet of modern CPSs is its holistic view of concerns, components, and implementation competences. The range of functionalities required of CPSs increasingly include Web-enabled components, which conjoins CPSs to the landscape of the next computing paradigm.

Figure 1: a cartoon-type representation of the conjoining of the elements listed above (created by Dall-E on 6 November 2023.

- **Digital Twins**, which are the digital representation of real-world entities, hosted on compute infrastructures that may or may not be digitally connected to their actual counterpart. Digital Twins have a tremendous potential, for science, learning, conceiving, building, optimizing, planning, maintenance. They are essential for the 4D computing because they allow to use time as a real variable: we can see what happed in the past by simulation, finding the cause of phenomenon's, and also to make forecast by analysing the evolution in the future, therefore finding better options. Some capture the potential of Digital Twins under the umbrella term of metaverses (this is a 3D rendering of the realm of computing of the digital twins, visualized for humans). Regardless of the denomination, it is easy to see that Digital Twins may be realized and exposed as web resources. If they are so, then they become part of the general (or specialized) Web space, thus making metaverses less secluded and self-contained and more permeable and pertinent to the vision we are discussing here.

Recently, two further important innovations have arisen, with potential to cause disruptive evolution of the landscape formed by the elements discussed above:

- **Generative AI**, a vertex of narrow (task-specific) artificial intelligence that is able to produce digital products of any sort, including computer programs and control commands, using "generative models". Such models recognize structural and correlative patterns in training data extracted from specific target domains, and return rich digital outputs that feature similar patterns consistent with the received prompts. Current-generation models are massive, for size of data and need of training, but there are also smaller models emerging that are usable in a particular context. These needs demand massive investments for their production, which makes them especially attractive to commercial exploitation. Most evidently, however, the disruptive power of innovation carried by such models is also of crucial public interest. This concern will likely promote two parallel routes of evolution: (1) the development and preservation of (regional, national, continental) open-source publicly regulated foundation models, i.e., the general-purpose platforms that support the creation of generative AI applications; (2) the development of task-specific learned models and associated engines that can be deployed on resource-constrained devices for personal use or in industrial or civil infrastructures.

Route (2) will push generative AI to the Edge, and have it render personalized services. In previous editions of the HiPEAC Vision we have discussed of such AI-assisted personalized services as Guardian Angel and Digel, short for Digital Genius Loci (Duranton, 2023). This section of the HiPEAC vision builds on the vision presented in those documents.

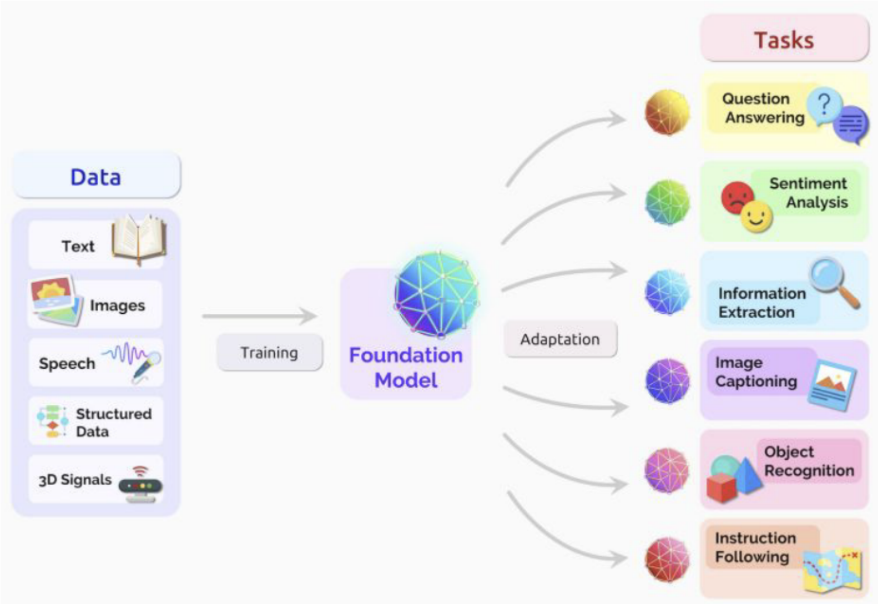

Figure 2: a view of Foundation Models as the next step from Large Language Models, and as enablers to task-specific models (source: "On the Opportunities and Risks of Foundation Models", Stanford University).

- The **Continuum of Computing**, as the digital integration of all elements listed above into a seamless networked platform where:
  - All available resources, regardless of their place of residence, are exposed as as-a-service web resources and accessed using streamlined Web-level protocols;
  - individual application services can be federated dynamically into ephemeral aggregations orchestrated into RESTful workflows originated at any point of the continuum, possibly constructed by task-specific edge-device-friendly generative AI engines;
  - the execution of the parts of those orchestrations will no longer be bound to a host device but may be able to move opportunistically from handheld or Edge devices to the centre of the Cloud, seeking to balance latency, privacy, data freshness, energy efficiency, and location requirements.

The latter prediction perhaps requires some rationale. It is vastly acknowledged that the Cloud-centric mode of service delivery causes all sorts and quantities of data to be transported from their source to the centre of the Cloud, where most application services reside, for specialized processing. This transfer is energetically costly, it is privacy-threatening, and it incurs delivery latencies that may be from fastidious to untenable for time-sensitive services. The natural countermeasure would be to have applications (or parts of them, including their orchestration logic) travel toward the data sources of interest, instead of the opposite. As value-added applications will likely be ephemeral and opportunistic, it is impractical that they be recompiled for their destination target and equally undesirable that their ability to travel be constrained by compatibility constraints. This observation carries the implication that the execution environment at any compute note of the continuum that wishes to be part of that infrastructure be able to host and execute these application components.

EVERYONE HAS THEIR OWN GUARDIAN ANGEL, WHO CAN PROTECT THEM,
SO THAT EVERYONE CAN MAKE THEIR WAY SAFELY THROUGH OUR DIGITAL WORLD.

Figure 3: a scene from the HiPEAC Comic Book published in 2019 (see: https://www.hipeac.net/media/public/files/46/7/HiPEAC-2019-Comic-Book.pdf), which evoked the Guardian Angel concept as it was described in [1].

Figure 4: How artificial intelligence (Dall-E 3) sees the continuum of computing

Key implications

The anticipated convergence of all the elements listed above will give rise to what the HiPEAC Vision 2024 terms "the **Next Computing Paradigm**", NCP. The key tenets of the NCP vision postulate:

- The integration of the "web of humans" with the "web of machines", where all the digital resources represented in that integration expose as-a-service interfaces that can be accessed, manipulated, and aggregated using Web-type protocols. To this end, such protocols will have to be maximally streamlined to become sustainable for use with all types of compute devices.
  - This direction of evolution will require the specification capabilities of the interface points for such Web-type protocols to be augmented to capture an increasing range of non-functional requirements (energy, latency, provenance, service level, etc.)
- The impetuous emergence of a spatial (and time) dimension to the next-generation web-inspired platform evoked above, which will be crucial to warrant context-awareness in the regard of physical constraints, location-dependent rules (e.g., norms and legislation), local knowledge. The spatial augmentation of Web-type protocols will:
  - Require a standard language to encode properties of physical objects and spaces, logical concepts and allowable activities associated with them.
  - Require a suite of standard protocols to expose contract-based interfaces associated to zones and objects, and to support credentialed requests and interrogations on them.
  - Equally apply to mobile computation (execution bundles that may move for and during execution) and to mobile devices (where executions are statically bound to a moving host), unmanned aerial vehicles, and other semi-autonomous transportation vehicles.

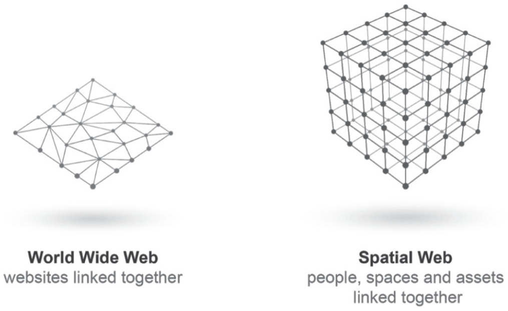

Figure 4: a view of the Spatial Web contrasted with the Web as it currently is. Excerpted from "The Spatial Web", by G. René and D. Mapes, 2019, page 35.

- The envisioned spatial computing will be CPS-like (operating with and for physical systems, coping with time constrains of the real world), swarm-like (supporting opportunistic dynamic and mobile aggregations of compute nodes within variable-size logical regions), 4D-enabled (fit for extended reality, spatial digital twins involved in time-sensitive operation).
- The embedding of Generative AI engines (models and prompt handlers) in Edge devices, whether handheld or deployed in industrial or civil installations to provide for the on-the-fly construction of ad-hoc applications expressed as dynamic, opportunities, ephemeral, smart orchestrations of calls to as-a-service interfaces exposed in the logical or physical regions of interest. Interestingly, such regions can well be temporary self-sufficient federated clusters of Edge devices that may even occasionally happen to be partitioned from the Internet. The prompts that will trigger creation, deployment, and execution of these dynamic orchestrations will use any conceivable "natural" interfaces, including voice for humans, and video imaging for automated requestors.

Conclusion

The document presents a transformative vision for the future of computing, emphasizing the integration of digital and physical realms through advanced web protocols, spatial awareness, and Generative AI. This paradigm shift aims to create a more efficient, context-aware, and user-centric digital ecosystem, paving the way for innovative applications and services in various sectors. The success of this vision hinges on technological advancements and collaborative efforts across multiple domains.

Figure 5: The Next Computing Paradigm, hallucinated by Dall-E 3.

References

Duranton, T. V. (2023). “Digels”, digital genius loci engines to guide and protect users in the “next web”. In M. D. al. (Ed.), *HiPEAC Vision 2023*, (pp. 18-21).

Tullio Vardanega is an associate professor in the Department of Mathematics at the University of Padova, Italy.

Marc Duranton is a researcher in the research and technology department at CEA (the French Atomic Energy Commission) and the coordinator of the HiPEAC Vision 2023.