Les composants pour les infrastructures numériques

Par Dimitri KTÉNAS et le Dr Denis DUTOIT
Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA)

Dans cet article, nous nous intéressons plus particulièrement aux composants constitutifs des infrastructures numériques, en premier lieu sous l'angle des infrastructures de calcul, puis sous l'angle des infrastructures de communication.

Les composants clés de l'infrastructure de calcul des systèmes numériques incluent les capteurs intelligents, les processeurs d'application, et tout xPU fournissant une capacité de calcul. Ces composants créent une infrastructure de traitement numérique hétérogène, flexible et évolutive, capable de répondre aux besoins croissants du calcul.

D'un autre côté, le domaine des télécommunications est devenu la clé de voûte de l'économie des pays occidentaux. Il apparaît que la 6G, en cours de définition, va nécessiter des composants dédiés nouveaux du fait notamment de la montée en fréquence qui risque de rendre inopérant les composants actuels en silicium. C'est donc une véritable rupture qui se prépare dans ce domaine de la microélectronique.

Introduction

L'évolution du monde numérique a toujours suivi celle des besoins de l'usager, de la facilité d'utilisation, des services offerts. Tout a commencé avec le besoin d'exécuter des calculs toujours plus nombreux et complexes, puis de rapprocher la source d'information au plus près de l'utilisateur. En parallèle sont apparus internet et le *smartphone*, lequel est aujourd'hui le moteur des avancées les plus importantes dans les technologies et les composants, y compris dans les nœuds technologiques très avancés.

Les évolutions liées à la transformation numérique sont supportées par des technologies sous-jacentes destinées à un déploiement massif telles que l'internet des objets, l'intelligence artificielle, la cybersécurité, le jumeau numérique. Avec nombre de défis à relever : qualité, sûreté, sécurité des services, gestion et traitement des flux massifs de données, miniaturisation de composants toujours plus performants, le tout avec des contraintes fortes de réduction de la consommation énergétique. Dans la suite de ce chapitre, nous nous intéressons plus particulièrement aux composants constitutifs des infrastructures numériques, en premier lieu sous l'angle des infrastructures de calcul, puis sous l'angle des infrastructures de communication. L'objectif poursuivi est de présenter les challenges à relever pour ces composants mais aussi de donner un aperçu de ce qu'ils pourraient être dans le futur.

Infrastructure de calcul des systèmes numériques

L'augmentation massive des données liées aux dispositifs IoT (*Internet-of-Things*), à la diffusion de contenus multimédias ainsi qu'à l'Intelligence Artificielle engendre une transformation majeure de l'infrastructure de calcul des systèmes numériques. Cette

évolution est bien détaillée dans les différentes feuilles de route du projet européen HiPEAC (https://www.hipeac.net/#/) depuis 2004. La Figure 1 résume cette évolution en trois phases :

- *Cloud computing* : initialement, le calcul et le stockage étaient principalement centralisés dans les centres de données traditionnels. Les capteurs du monde physique remontaient les données brutes vers ces centres pour y être traitées.
- Systèmes cyber-physiques intelligents : avec l'augmentation des dispositifs IoT et une volonté de gagner en efficacité énergétique, l'architecture a évolué pour inclure des composants de traitement numérique plus proches de la source de données. Ainsi, les microcontrôleurs et les capteurs IoT sont devenus plus sophistiqués, capables de prétraiter les données avant de les envoyer à des passerelles (*gateway*) ou des serveurs à la périphérie (*edge computing*). Ces serveurs ont gagné en puissance de calcul et sont maintenant capables de traiter des applications d'inférence en intelligence artificielle, réduisant ainsi le besoin de transfert de données volumineuses vers le *cloud* (F. Palumbo *et al.*, 2023).
- Systèmes distribués coopératifs : les besoins applicatifs à la périphérie vont encore évoluer vers plus de puissance de calcul, plus de données à traiter, le tout avec une faible latence. L'exemple le plus probant étant l'intelligence artificielle générative. Les capacités de calcul à la périphérie ne seront plus suffisantes, notamment pour l'apprentissage incrémental de ces modèles. C'est pourquoi une nouvelle couche de petits serveurs distribués et coopérant entre eux, situés plus près des utilisateurs finaux, va se créer pour offrir une puissance de calcul intermédiaire. Ainsi, les ressources de stockage et de traitement seront réparties entre le *cloud* centralisé dans les centres de données, le *cloud* distribué, les serveurs à la périphérie et enfin les capacités de calcul des capteurs pour former le continuum de calcul ou continuum informatique. Ce modèle permet d'optimiser la gestion des données et les processus de calcul en fonction de leur emplacement et de leurs exigences spécifiques.

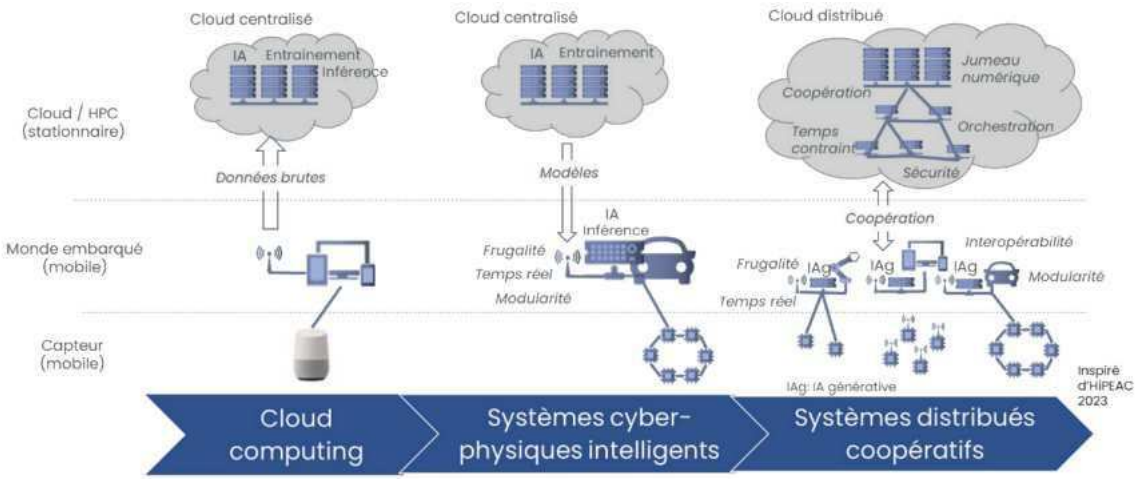

Figure 1 : Évolution du continuum informatique (Source : HiPEAC 2023).

Les composants clés de ce continuum incluent :

- les capteurs intelligents pour la couche de calcul proche du monde physique et ainsi collecter et prétraiter les données localement pour réduire la consommation de bande passante et d'énergie ;
- les processeurs d'application des téléphones mobiles ou encore les processeurs dits de fusion pour l'électronique de l'automobile offrent une puissance de calcul intermé-

diaire au niveau de la périphérie pour fusionner et traiter des données provenant des capteurs ;

- enfin, les CPU, GPU, TPU ou tout autre xPU fournissent une capacité de calcul à grande échelle dans le *cloud* centralisé et à plus petite échelle dans le *cloud* distribué plus proche des utilisateurs.

Ensemble, ces composants créent une infrastructure de traitement numérique hétérogène, flexible et évolutive, capable de répondre aux besoins croissants du calcul. L'évolution des composants dépend de leur position dans le continuum de calcul, comme expliqué ci-après.

Pour les composants proches du monde physique (capteurs), le principal challenge réside dans l'atteinte de l'ultra-basse consommation. Les composants devront ainsi être de plus en plus autonomes au niveau énergétique, avec des architectures en rupture incluant le calcul directement au sein des capteurs *via* une intégration multicouches 3D (*i.e.* un assemblage vertical de puces) hétérogène. En d'autres termes, l'architecture du composant incorporera une couche capteur, une couche pré-traitement, une couche mémoire et une couche de traitement de l'Intelligence Artificielle (IA). Ce type de solution est déjà utilisée, par exemple dans le cas des capteurs d'images pour la téléphonie mobile. La France est bien située dans le domaine de l'intégration 3D, que ce soit au niveau de la recherche avec l'IRT Nanoélectronique ou au niveau industriel avec les imageurs de ST Microelectronics ou encore de la société Prophesee, dans ce que l'on appelle la vallée des capteurs d'images aux alentours de Grenoble. Dans un autre registre, une société comme Greenwaves propose des processeurs d'application très basse consommation pour les objets connectés.

Pour les composants du *cloud*, le challenge a basculé ces dernières années sur le traitement de l'inférence et de l'apprentissage des modèles d'IA génératives. En 2022, un modèle d'IA générative (par exemple GPT-3) comporte déjà 200 milliards de paramètres alors que la mémoire de travail d'un GPU approche seulement les quelques dizaines de milliards d'octets. Et cet écart va en s'accroissant. La mémoire et son utilisation efficace en énergie deviennent désormais le principal challenge de ce segment du continuum. Les mémoires de type "High Bandwidth Memory" utilisées pour ces applications d'intelligence artificielle se feront de plus en plus denses avec l'assemblage en vertical de plus de 12 couches de puces mémoires et une bande passante avec le GPU de plus en plus élevée comme les futures HBM4. Les GPU deviennent les seuls composants capables de traiter les modèles d'IA générative dans le *cloud* avec une domination de NVIDIA et son dernier Blackwell, fabriqué en 4 nm et qui contient plus de 200 milliards de transistors. AMD avec son MI300 se positionne comme le principal concurrent de NVIDIA. Pour encore augmenter les capacités de calcul, une société californienne récente, Cerebras, utilise une tranche de silicium complète pour son processeur (wafer-level computing). C'est donc la course au gigantisme à laquelle la France n'est pas du tout présente. Pour le *cloud*, la France n'apparaît en effet que pour les CPU avec l'Initiative de Processeur Européen (EPI) d'EuroHPC et la conception du processeur Rhea de la toute récente société SiPearl. Une autre évolution majeure du *cloud* réside dans l'intégration de la photonique dans les communications entre les racks des serveurs. Cette communication optique, arrive désormais entre les boitiers des composants avec des circuits d'entrées/sorties électro-optiques pour gagner en latence de transmission et ainsi accroître encore la capacité mémoire et de traitement d'un nœud de calcul. Enfin devrait arriver d'ici une dizaine d'années le calcul analogique optique où les multiplications-accumulations ne seront plus numériques mais optiques. Cela permettra de franchir plusieurs ordres de grandeurs en capacité de calcul et en efficacité énergétique. Mais de nombreux challenges restent à relever dans ce domaine comme l'énergie nécessaire pour les conversions électro-optiques inévitables (les calculs non linéaires ne peuvent pas se faire en optique) ou la taille des filtres optiques mis en œuvre pour le calcul.

Pour l'embarqué, les performances de calcul et les capacités de mémoire sont d'un à deux ordres de grandeurs moindres que pour le *cloud*. Pour garder une bonne efficacité énergétique, des architectures de rupture seront tout de même nécessaires comme le calcul proche/dans la mémoire. Beaucoup de société *fabless* se créent autour de ce segment de l'Intelligence Artificielle en périphérie (embarquée) et aussi en France avec Kalray, VSORA, Menta et d'autres encore. ST Microelectronics se positionne également sur ce segment avec ses microcontrôleurs pour le marché de l'automobile. C'est dans l'embarqué, pour l'inférence en IA que le calcul analogique a retrouvé un regain d'intérêt avec plus particulièrement les circuits neuromorphiques qui tentent d'émuler un réseau de neurones en remplaçant les multiplications-accumulations numériques par un maillage de résistances où les courants s'additionnent. Le challenge de ces architectures analogiques est d'avoir un réseau de résistances suffisamment dense pour émuler un réseau complet. Pour cela, les nouvelles mémoires résistives non-volatiles sont prometteuses.

Si l'on se projette plus avant, à l'horizon 2040, on peut raisonnablement anticiper une interconnexion accrue des systèmes informatiques, associée à l'intégration de nouvelles technologies et architectures dans les composants. Cette évolution augmentera les capacités de calcul et l'efficacité énergétique, tout en permettant une granularité plus fine des systèmes distribués. Cependant, cette évolution n'est pas sans risque. Les travaux de C. Weir *et al.* (2024) identifient quelques impacts négatifs de cette évolution comme la difficulté de distinguer la réalité de la fiction due à la prolifération de la désinformation générée par l'IA ou encore la difficulté de comprendre le comportement des systèmes et ainsi de différencier une attaque intentionnelle d'un incident mineur. Pour atténuer ces risques, les auteurs préconisent quelques recommandations comme la « responsabilité ambiante » ou plus précisément la capacité des systèmes à se vérifier eux-mêmes pour garantir la sécurité ou encore l'investissement dans des méthodes de développement responsable avec de meilleures pratiques en ingénierie logicielle pour les systèmes d'IA.

Infrastructure de communication des systèmes numériques

Le domaine des télécommunications, et plus particulièrement des communications sans-fil, est aujourd'hui devenu la clé de voute de l'économie des pays occidentaux. Depuis l'avènement des *smartphones*, c'est également un domaine structurant pour la microélectronique. Le déploiement de la 5G met en évidence l'importance pour notre souveraineté de maîtriser les technologies de communication. Les batailles technologiques et une partie des tensions géopolitiques d'aujourd'hui sont nées autour des technologies de communication (3G, 4G, 5G et demain 6G). La France garde un historique fort sur ce domaine malgré l'évolution des grands acteurs (acquisition d'Alcatel par Nokia, opérateur historique se tournant vers le service) et l'Europe héberge toujours des grands industriels (Nokia, Ericsson) du secteur (principalement l'infrastructure).Tout comme les précédentes générations de systèmes de télécommunication, le déploiement de la 5G est rendu possible grâce aux avancées de la microélectronique (Loi de Moore), qui permet un traitement de l'information plus efficace énergétiquement. Les gains visibles pour les utilisateurs sont une connectivité plus facile amenée par des débits plus élevés et des temps de latence plus faibles. À un moindre degré, une meilleure couverture est également attendue. Aussi, ces technologies de communication ont largement dopé le marché des semi-conducteurs.

Des travaux sont d'ores et déjà en cours pour définir le périmètre et les spécifications de la 6G. Cette génération devrait arriver sur le marché à l'horizon 2030. D'emblée, il apparaît que cette génération va nécessiter des composants dédiés nouveaux du fait notamment de la montée en fréquence qui risque de rendre inopérant les composants actuels en silicium. C'est une véritable rupture qui se prépare donc dès maintenant dans ce domaine de la microélectronique avec l'utilisation de matériaux nouveaux (introduc-

tion de matériaux III-V) plus adaptés aux fréquences hautes. Plus précisément, la 6G va s'accompagner d'une montée importante en fréquence des bandes de communication, permettant plus de bande passante, de débit et une latence réduite, depuis les bandes millimétriques (> 12 GHz) vers les bandes sub-THz. Cette montée en fréquence changera drastiquement la propagation des ondes et *in fine* les réseaux puisque même en considérant des communications directives, par opposition aux antennes omnidirectionnelles ou à la segmentation de l'espace actuels, les portées seront plus faibles à ces fréquences avec donc beaucoup plus de stations de base envisagées.

En parallèle une autre révolution s'opère dans les infrastructures télécoms avec le déploiement de nouveaux Radio Access Networks (voir la Figure 2) : entre les « têtes RF » / RRU – Remote Radio Unit – délivrant et recevant les signaux, et le cœur de réseaux fournissant les données / CN – Core Network, l'architecture est devenue entièrement reconfigurable et programmable (virtualization-RAN), répartie entre Distributed Unit – DU – très nombreuses et Centralized Unit – CU – regroupant les précédentes.

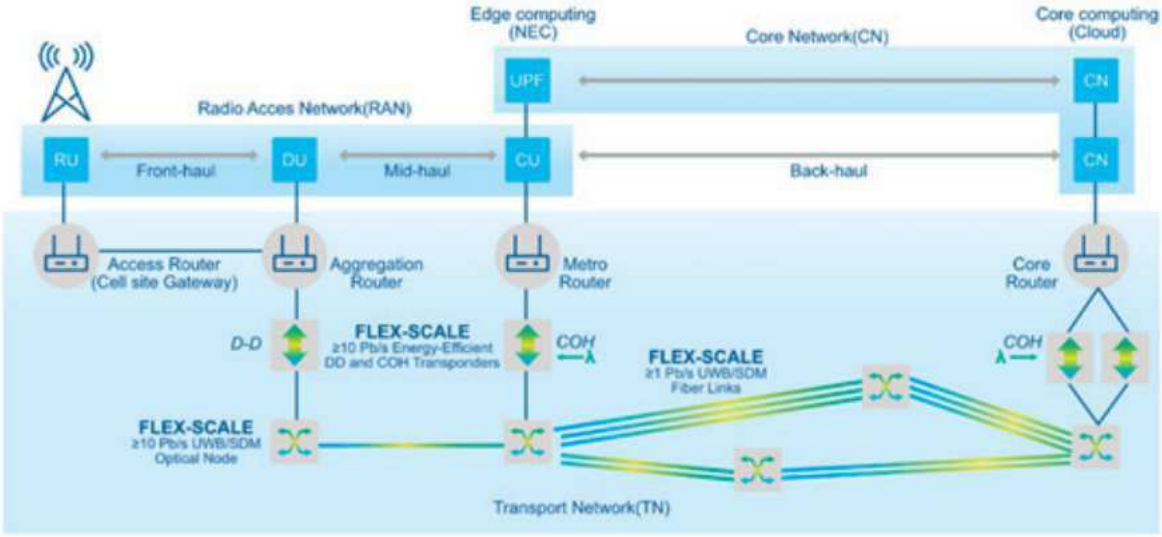

Figure 2 : Réseau de bout-en-bout vu par l'ETSI
pour le B5G ou 5.5G (Source : Effenberger, 2022).

Cette reconfigurabilité et programmabilité implique que les acteurs fournisseurs de ces solutions changent et que seront favorisés les fournisseurs de solutions matérielles et logicielles les plus performantes. Il sera alors plus facile d'insérer ces RRU fonctionnant en bandes hautes au sein d'un réseau offrant plus d'opportunités de découpage et de virtualisation, avec à chaque niveau la possibilité de mise en œuvre d'outils logiciels qui favoriseraient les acteurs américains ou japonais, dans un premier temps, déjà bien connus au détriment des fournisseurs d'infrastructures dédiées télécom tels que Nokia et Ericsson. La nouvelle architecture portée par l'Open RAN nécessitera d'avoir des processeurs au niveau des DU et des CU. Une première indication semble favoriser un acteur tel que Nvidia qui est *leader* dans le domaine des processeurs graphiques, très bien adaptés aux besoins des réseaux, et conçus dans des technologies FinFet très fines. En Europe, un acteur montant comme Kalray pourrait jouer un rôle essentiel dans l'avènement d'une solution souveraine, utilisant des technologies continentales.

Concernant la fourniture de composants pour les télécoms, les acteurs européens n'ont qu'une petite part face à des concurrents tels que Qualcomm, MediaTek, NVidia, Apple, Intel/Altera, AMD/Xilinx… qui sont au premier plan pour les terminaux, mais également pour les réseaux. Les acteurs européens majeurs – STMicroelectronics, NXP, Infineon –

sont en revanche *leaders* pour les composants en lien avec les marchés verticaux associés aux applications industrielles.

De leurs côtés, STMicroelectronics et GF proposent des technologies adéquates pour des composants RF mais également des processeurs adaptés à certaines applications : FD-SOI 28 nm et 22 nm, avec évolution vers le 18 nm puis le 10 nm, FinFet 12 nm pour GF, BiCMOS 55 nm pour ST. Ces technologies fines proposées ont déjà démontré leurs intérêts pour des applications RF, c'est-à-dire au niveau des RRU dans les réseaux, pour la connectivité *wifi* et pour les circuits bas coût de l'IoT. À l'inverse, ils ne disposent pas de finesse de gravure suffisante – 5 nm et en dessous – pour entrer dans les terminaux mobiles. Les enjeux pour les acteurs français et européens résident dans leur capacité à répondre au besoin en approvisionnement de composants sur des parties bien déterminées du réseau.

Dans les RRU, les têtes RF seront de différents types, en fonction de la fréquence et des performances demandées. Des industriels européens tels que STMicroelectronics peuvent proposer des technologies hybrides basées sur le B9MW (130 nm BiCMOS) et le B55X (55 nm BiCMOS) et co-intégrant des technologies de type GaN pour améliorer les performances en efficacité énergétique en sub-7 GHz et bandes K, et de type InP pour la montée en fréquence pour les bandes mmW 60 GHz, 140 GHz voire 300 GHz (Calvanese *et al.*, 2021). De même, pour les réseaux, l'Europe a une carte à jouer dans le domaine de l'optique et notamment de la photonique sur silicium. Les réseaux télécoms vont être massivement déployés autour de réseaux de fibres optiques dont les composants de tête – lasers et diodes – co-intègrent des technologies II-VI pour l'optique et CMOS/BiCMOS pour les parties électroniques. Ce type de co-intégration correspond parfaitement au modèle des fonderies européennes, qui peuvent fournir les technologies de base et approvisionner celles manquantes pour assurer l'assemblage.

Conclusion

L'émergence de cœurs de réseaux virtualisés va apporter de la flexibilité et permettre la connexion de nombreuses modalités de communication sans fil autour d'un traitement numérique très puissant et flexible. Une complexité globale accrue, et en corollaire, une maîtrise difficile de l'ensemble des données échangées font apparaître une difficulté en matière de souveraineté et la possibilité de captation d'informations. La tendance actuelle de rachat par les Gafam, ou d'alliances stratégiques dans le monde des Telecom, ne laisse aucun doute sur l'importance de la maîtrise de ce cœur de réseau et de la rupture à accomplir en Europe pour rattraper ces acteurs mondiaux américains. Par-dessus cette maîtrise du monde numérique, il faut aussi considérer l'importance de la gradation des traitements opérés : du *mobile edge computing* proche de la source d'information au *cloud* générique et hyper-concentré, différents niveaux vont nécessiter d'être développés et maîtrisés.

Autour de ce réseau central, la variété des modalités RF va multiplier les systèmes radio déployés : plus de composants, de matériel et d'objets technologiques disséminés. Il conviendra donc d'être attentif à la neutralité écologique de ces systèmes, à l'acceptation sociétale et en regard des besoins accrus en composants, être assuré de la viabilité de la filière microélectronique européenne pour réduire la dépendance actuelle à la Chine, à Taïwan et aux États-Unis. Des telles technologies existent déjà pour répondre à une partie des besoins, notamment les technologies à base de SOI (FD-SOI ou RF-SOI), mais l'évolution rapide des communications sans-fil, et l'accroissement de connectivité dans les applications va nécessiter de considérer l'élargissement des solutions actuelles vers plus de diversité, en particulier vers des nœuds technologiques plus avancés.

Bibliographie

PALUMBO F., LAZCANO R. & MADRONAL D. (2023), "Towards a living dimension: The future of cyberphysical systems", *in* DURANTON M. *et al.*, editors, *HiPEAC Vision 2023*, pp. 44-53.

WEIR C., DYSON A., JOGUNOLA O., DENNIS L. & PAXTON-FEAR K. (2024), "Interlinked computing in 2040: Safety, truth, ownership, and accountability", *Computer*, vol. 57, n°1, pp. 59-68.

EFFENBERGER F.J. (2022), "Fixed 5th generation advanced and beyond", 1st edition – September, ETSI White Paper N°#50, ISBN N°979108262071.

CALVANESE STRINATI E., BELOT D., FALEMPIN F. & DORE J.B. (2021), "Towards 6G: from new hardware design to wireless semantic and goal-oriented communication paradigms", ESSCIRC 2021, 6-20 September 2021.