Allocation de fréquences pour les Interconnexions RF dans un Réseau sur Puce Multi-Cœurs Massivement Parallèle

Eren UNLU, Christophe MOY

CentraleSupélec/IETR
Avenue de la Boulaie, 35576 Cesson Sévigné, France
eren.unlu@centralesupelec.fr, christophe.moy@centralesupelec.fr

Résumé – Des solutions radio-fréquences (RF) et optiques ont été proposes ces dernières années pour répondre aux besoins en termes de bande passante des réseaux sur puce massivement parallèles. Cependant, ces approches ont souvent été considérée inacceptables en raison de la complexité de la mise en œuvre de nombreux éléments supplémentaires en technologie CMOS d'une part, et de leur manque de flexibilité d'autre part. Afin de contrecarrer ces limitations, nous proposons d'utiliser l'OFDMA (accès multiple par division de fréquences orthogonales) comme support d'une approche flexible d'allocation des ressources RF pour l'interconnexion des éléments de traitements. Dans cet article, nous proposons un ré-arbitrage de la bande passante RF en fonction du trafic instantané entre les éléments du réseau sur puce. Un protocole d'ordonnancement proportionnel à la taille attendue des files d'attente (Expected Queue Proportional Scheduling - EQPS) est présenté pour ce contexte OFDMA. Il permet d'atteindre une réduction du retard moyen et instantané de la date de livraison des données pouvant atteindre un facteur 7,5 tout en assurant une équité entre les éléments du réseau. Les contraintes d'implantation ne sont pas dans le champs de cet article qui se concentre sur le problème de l'allocations des ressources de communications.

Abstract – Radio Frequency (RF) and optical interconnects have been introduced recently in many-core Network on Chip (NoC) in order to satisfy the high on-chip bandwidth demand. However, the necessity to implant huge number of CMOS instruments has been considered prohibitive and the solutions proposed until now suffer from a lack of flexibility due to their static nature. In order to overcome these drawbacks, we propose Orthogonal Frequency Division Multiple Access (OFDMA) as an effective flexible allocation scheme for an RF interconnect. This paper shows how we propose to re-arbitrate the bandwidth among on-chip elements according to instantaneous traffic. An effective distributed dynamic bandwidth allocation protocol, called Expected Queue Proportional Scheduling (EQPS) is provided for this OFDMA interconnect, which is shown to provide up to x7.5 lower average and instantaneous latency while ensuring fairness among nodes. Implementation matters are out of the scope of this paper, which focuses on the communication resource allocation issue.

1 Introduction

En réponse aux applications numériques de plus en plus complexes et rapides, les progrès de l'industrie de la micro-électronique permettent d'envisager désormais la parallélisation de milliers de cœurs de traitements dans une même puce. On parle d'architectures massivement parallèle ou CMP (Chip Multi-Processors). La répartition de la mémoire physique entre les cœurs impose une approche de gestion de mémoire partagée et distribuée avec cohérence de cache. Des protocoles de cohérence de cache permettent alors de considérer tous ces espaces mémoire physiques différents, comme une seule mémoire globale au circuit, au niveau logique. Les nombreuses communications engendrées par ces protocoles créent un nouveau goulot d'étranglement au sein de ces architectures massivement parallèle, dont la complexité a par conséquent connu une translation du calcul vers les communications [1]. On parle alors de réseau sur puce ou NoC (Network on Chip). Les réseaux sur puce conventionnels sont un maillage câblé matriciel à deux dimensions (2D) ou chaque cœur de traitement est lié par un routeur au réseau en quadrillage [2]. Mais des limites de congestion du réseau sont atteintes pour quelques centaines de cœurs, en termes de délai de transmission, d'incapacité à faire de la diffusion efficacement ou même de blocage [3]. C'est pourquoi ont émergé ces dernières années de nouvelles approches à base de transmission optiques [3] ou radiofréquence (RF) [4], permettant de transmettre plusieurs flux simultanés sur des bandes de fréquences différentes. Cependant, les approches proposées jusqu'ici ne permettaient pas un changement dynamique d'allocation des ressources de communication en fonction des besoins de transmission en temps-réel. Or comme le trafic de cohérence de cache est hautement hétérogène à la fois temporellement et spatialement, les besoins instantanés de transmission des cœurs sont très variables [5].

Afin de contrecarrer ces inconvénients, nous proposons, dans le cadre du projet ANR WiNoCoD [6], de spécifier et réaliser en technologie CMOS standard, un NoC RF guidé, basé sur un accès multiple orthogonal par division de fréquence (OFDMA) [7]. Nous n'allons pas détailler ici les principes de l'OFDMA et de l'OFDM sur laquelle elle repose. Pour notre cas d'étude, il suffit de savoir notamment que l'OFDMA permet nativement de mettre en œuvre des procédés de diffusion ainsi que

des schémas d'allocation dynamique de la bande passante de communication. Pour ce dernier aspect, il suffit simplement à l'émission, d'organiser les données en vecteurs de N éléments et de les ordonner en entrée du modulateur, constitué notamment d'une IFFT (Inverse Fast Fourier Transform) de taille N, afin que la ième donnée du vecteur soit transmises sur la ième sous-porteuse ou fréquence du peigne OFDM. Ainsi pour changer l'allocation de fréquence, il suffit de changer l'ordre des données dans le vecteur d'entrée. Si plusieurs émetteurs (synchronisés) veulent émettre chacun sur des sous-porteuses OFDM différentes d'un même symbole, il leur suffit de choisir de positionner leurs symboles dans leur vecteur d'entrée à des positions différentes et de mettre un zéro dans toutes les autres positions du vecteur. Ainsi à chaque émetteur est alloué un jeu de sous-porteuses exclusif des autres émetteurs, et la fréquence des sous-porteuses identifie l'émetteur en mode «écriture simple, lecture multiple» ou SWMR (Single Write, Multiple Read).

En ce qui concerne les capacités de diffusion, tout récepteur qui effectue une FFT de taille N sur ce les symboles reçus est en capacité de démoduler toutes les sous-porteuses et sélectionner celles qui le concerne par analyse de l'identifiant du destinataire, un identifiant particulier pouvant être attribué pour le mode diffusion afin que tous puissent le lire.

2 Architecture WiNoCoD

2.1 Architecture de la puce multi-processeurs

L'architecture de la puce servant de référence au projet WiNoCoD est présentée sur la. Les cœurs de traitement sont regroupés, sur la droite de la figure, en tuiles (tiles) de 4 cœurs associées à un cache d'instruction et de données de niveau 1 implémentant un protocole HDBCC (Hybrid Directory Based Cache Coherence) [3][8], une mémoire RAM et un DMA, tous mis en connexion entre eux par un crossbar. Les tuiles sont regroupées par 16 pour former une grappe (cluster), comme illustré au centre de la . Les tuiles de la grappe sont connectées par un maillage 2D se rattachant au crossbar de chaque tuile. Comme on peut le voir à gauche de la , un lien radio guidé, en forme de serpentin, permet aux 32 grappes de communiquer entre elles via une interface d'émission/réception OFDMA nommée RF NoC Interface. Ainsi le nombre d'accès RF dans la puce, et donc de modulateurs/démodulateurs RF qui vont occuper de la surface et consommer de l'énergie est limité au nombre de grappes. Ces aspects sont hors du sujet de cet article, mais peuvent être obtenus dans la référence [9].

2.2 Paramètres radio du RF NoC

Les principales caractéristiques de transmission du RF NoC sont les suivantes : (*i*). Bande passante de 20 GHz, centrée sur 30 GHz et durée d'un symbole OFDM de 50 ns [8], (*ii*). 1024 sous-porteuses et les sous-porteuses sont groupées par blocs de 32 éléments qui représentent une unité atomique de ressources

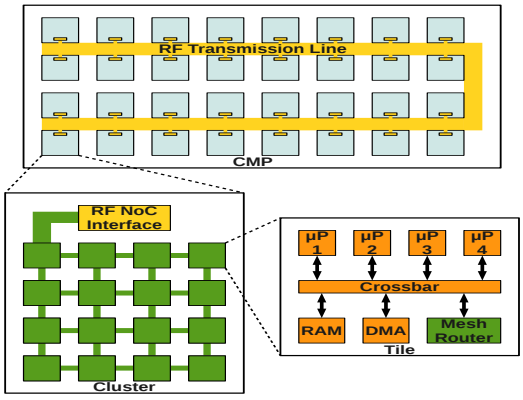

FIGURE 1 – Structure hiérarchique de l'architecture proposée à 2048 cœurs de traitement.

de communication, appelée «resource block» (RB) : Ainsi il y a un RB par grappe en moyenne.

En modulation QPSK (par défaut mais modifiable), soit 2 bits/sous-porteuse par symbole OFDM, cela permet d'atteindre un débit global maximal sur le RF NoC de 40 Gbps. Cette valeur n'est pas exceptionnelle, comparée aux plus de 1 Tbps obtenus par [3] en optique. D'une part nous avons restreint la bande passante afin de rester dans des limites réalisables aujourd'hui en technologie CMOS pour tous les éléments des chaînes RF. Une extrapolation peut être faite sur les futures technologies pour obtenir de bien meilleures performances. Mais d'autre part, ce que nous souhaitons mettre en valeur n'est pas le débit total lui-même, mais la capacité à utiliser tout ce débit à chaque instant, grâce à l'allocation flexible des sous-porteuses, à la demande, entre les grappes. Ceci représente un défi d'autant plus important que nous souhaitons le faire quelles que soient les demandes d'accès au RF NoC, i.e. homogènes à toutes les grappes, ou très hétérogènes notamment lorsque des grappes sont inutilisées.

Ainsi, un mode par défaut de fonctionnement du RF-NoC est le suivant, au niveau radio : chaque grappe, numérotée de 1 à 32, «possède» pour émettre, en SWMR, le bloc de ressource de même indice, sur lequel il peut transmettre 64 bits en QPSK, soit la taille d'un message unitaire ou flit (flow control digits). Ce mode par défaut est efficace si toutes les grappes exigent d'accéder au RF NoC de manière équilibrée en moyenne, afin de transmettre des données vers d'autres grappes. En revanche, si les demandes d'accès sont déséquilibrées, des délais peuvent apparaitre pour les données des grappes ayant un nombre de transmissions provoquant une saturation de l'occupation de leurs ressources, pendant que d'autres grappes n'utilisent pas les leurs. Notre cas d'étude équivaut donc à un problème où K=32 files d'attente doivent écouler leur flux à travers une ressource partagée. Nous proposons d'étudier ici un algorithme permettant de réduire à la fois la latence moyenne et le retard maximal de transmission des données à travers le RF NoC, et la probabilité de dépassement de la taille des files d'attente d'émission. Cela revient à appliquer les concepts de la radio intelligente à l'intérieur d'une puce [10].

3 Algorithme d'Allocation Dynamique de la bande passante

3.1 Contexte spécifique

Nous proposons un nouvel algorithme EQPS (Expected Queue Proportionnal Scheduler). C'est une version modifiée de l'algorithme QPS dont il est démontré qu'il offre des délais de transmission faibles, tout en assurant une égalité de traitements entres les files d'attente, lorsque le système réagit instantanément à l'arrivée de paquets dans les files [11]. Dans notre cas, le système ne pourra ré-allouer les RB pour donner accès au RF NoC aux grappes qu'avec une période de T symboles OFDM, pendant laquelle les files d'attente pourront avoir reçu des données à transmettre. Nous estimons en effet que les 50 ns séparant deux symboles OFDM ne seront pas suffisants pour effectuer une allocation coordonnée entre toutes les grappes qui minimise le surcoût de messagerie entre les grappes.

Le principe de coordination décentralisée entre les grappes est le suivant : grâce aux propriétés intrinsèques de diffusion de l'OFDMA, toutes les grappes vont diffuser à toutes les autres grappes, en début de période de T symboles et sur des sous-porteuses pré-définies, une information sur l'état de remplissage de leur file d'attente (Queue State Information ou QSI). Toutes les grappes reçoivent ces informations et exécutent le même algorithme d'allocation qui va permettre de décider de l'allocation des porteuses entre les grappes pour la période suivante, sans conflit. Il est à noter que ces informations de QSI ne consomment que peu de ressources : 8 bits par grappe pour considérer 256 niveaux, soit 128 sous-porteuses sur 1024. Ce schéma est illustré en où l'on voit que le surcoût de messagerie pour l'envoi du QSI n'est qu'un faible nombre de porteuses d'un symbole OFDM tous les T symboles. Bien sûr il existe un compromis entre T, et ce surcoût. Augmenter T consiste à diminuer ce surcoût, mais également à dégrader les performances de l'algorithme EQPS par rapport aux résultats idéaux de l'algorithme instantané QPS. Cependant cette discussion est

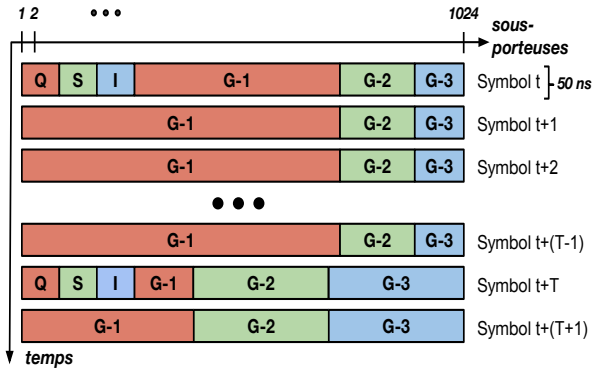

FIGURE 2 – Fonctionnement de l'algorithme EQPS permettant une allocation des sous-porteuses par rotation temporelle (de haut en bas) et fréquentielle (de gauche à droite) en fonction du QSI.

en-dehors du champ du présent article et peut être trouvée dans .

3.2 Algorithme EQPS

L'algorithme EQPS cherche à anticiper la valeur des QSI des grappes pour la prochaine période t+T, en termes de nombres de flits. Le «QSI anticipé» (expected QSI) d'une grappe dans une période donnée est fonction du QSI à la période précédente Q_i^t, de l'allocation en termes de nombre de RB (donc de sous-porteuses) dont cette grappe bénéficie à la période présente S_i^t, et du nombre de nouveaux flits attendus en moyenne pendant une période dans la file A_i^t :

$$\hat{Q}_i^{t+T} = max(0, Q_i^t - S_i^t) + A_i^t \quad (1)$$

Le calcul de la moyenne $\hat{A}_i^t$ est glissant, basé sur un coefficient α qui peut être par exemple de 0.95 :

$$\hat{A}_i^{t+T} = \alpha \hat{A}_i^t + (1-\alpha)\hat{X}_i^t \quad (2)$$

Le nombre de sous-porteuses allouées à la grappe i au symbole t+T est :

$$N_i^{t+T} = \left\lceil \frac{N\hat{Q}_i^{t+T}}{\sum_{\forall j} \hat{Q}_j^{t+T}} \right\rceil \quad (3)$$

Nous pouvons voir sur la Figure 2 un exemple simplifié avec 3 grappes. Leurs QSI sont diffusés toutes les périodes T et sur chaque période, une répartition des RB est effectuée entre les grappes, sur la base des QSI envoyés lors de la période précédente. Le calcul des la répartition est effectué dans chaque grappe selon l'équation (3). La répétition de ce calcul est préférable à la centralisation puis l'envoi du résultat pour chaque grappe à travers le NoC puisqu'il est très simple.

3.3 Résultats de simulation

Nous utilisons OMNET++, un simulateur à événements discrets, pour évaluer les performances de notre algorithme. Pour les simulations, il est très important d'émuler sur le RF NoC un trafic représentatif d'une application réelle impliquant une forte proportion de messagerie de cohérence de cache, notamment pour des actions de lecture, écriture, confirmation de réception. Nous avons tiré de [12] un modèle de flux des communications qui existe entre les cœurs d'un CMP, composé à 75% de paquets couts de 1 flit et 25% de paquets longs de 9 flits (576 bits). Au lieu de générer un trafic peu réaliste de type uniforme issu de chaque grappe, nous nous sommes basés sur les observations effectuées dans [12] sur le trafic dans les CMP.

La Figure 3 compare, lorsque les grappes n'injectent pas la même quantité de données dans le RF-NoC, les résultats de l'algorithme EQPS proposé avec celui d'un partage égal des ressources (Equal Share) tel que proposé dans les solutions optiques ou radio sans allocation dynamique des ressources. La distribution de l'injection des grappes est gaussienne entre les 32 grappes, comme dans [5] avec une moyenne d'injection de

1 paquet (soit 192 bits en moyenne) par symbole OFDM soit moins de 10% des capacités du RF NoC.

La Figure 3 montre en ordonnée la probabilité qu'un paquet excède un certain délai donné en abscisse. A titre d'exemple la probabilité qu'un paquet ait un délai d'envoi de 20 symboles OFDM est inférieure à 10-5, alors qu'elle est supérieure à 10% dans le cas d'une répartition statique et égale entre toutes les grappes. Une probabilité équivalente de livraison dans le cas statique est obtenue pour un délai de presque 160 paquets. On peut voir ainsi que l'algorithme EQPS proposé présente des résultats bien supérieurs à l'état de l'art des NoC optiques ou RF NoC sans allocation dynamique des ressources de communication, lorsque la charge sur le NoC est déséquilibrée entre les grappes. C'est un contexte qui se produira à chaque fois que l'ensemble des cœurs du CMP ne sera pas utilisé, ou même tant que le trafic entre les cœurs ne sera pas pris en compte à la compilation pour allouer le code sur les cœurs.

4 Conclusion

Ce papier propose un schéma de communication original et flexible basé sur de l'OFDMA pour utiliser des RF NoC dans les architectures de traitement massivement parallèles. Il est complété par d'autres publications qui expliquent pourquoi cette solution est peu couteuse en surface et consommation au regard du bénéfice offert en termes de souplesse et facilité de mise en oeuvre de solutions d'allocation flexible des ressources de communications sur le RF-NoC. L'algorithme EQPS proposé ici est un exemple permettant de minimiser le retard d'envoi des paquets dans le RF NoC dans des conditions réalistes de trafic.

Remerciements

Ce travail est supporté par l'ANR, dans le cade du projet WiNoCoD No : ANR-GUI-AAP-05. Les auteurs tiennent aussi à remercier les partenaires du projet : ETIS, LIP6 et NXP.

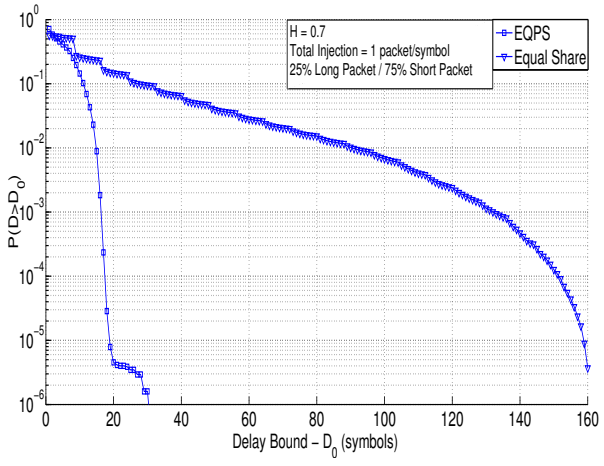

FIGURE 3 – Latence instantanée de l'algorithme EQPS proposé par rapport au cas de référence à partage égal des ressources, dans un cas de trafic réaliste sur le RF NoC d'un CMP

Références

[1] S. Pasricha and N. Dutt, *On-chip communication architectures : system on chip interconnect*. Morgan Kaufmann, 2010.

[2] W. J. Dally and B. Towles, "Route packets, not wires : On-chip interconnection networks," in *Design Automation Conference, 2001. Proceedings*. IEEE, 2001, pp. 684–689.

[3] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal, "Atac : a 1000-core cache-coherent processor with on-chip optical network," in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. ACM, 2010, pp. 477–488.

[4] M.-C. F. Chang, E. Socher, S.-W. Tam, J. Cong, and G. Reinman, "Rf interconnects for communications on-chip," in *Proceedings of the 2008 international symposium on Physical design*. ACM, 2008, pp. 78–83.

[5] V. Soteriou, H. Wang, and L.-S. Peh, "A statistical traffic model for on-chip interconnection networks," in *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. MASCOTS 2006. 14th IEEE International Symposium on*. IEEE, 2006, pp. 104–116.

[6] A. Briere, J. Denoulet, A. Pinna, B. Granado, F. Pêcheux, P. Garda, M. Ariaudo, F. Drillet, C. Duperrier, M. Hamieh *et al.*, "Winocod : Un réseau d'interconnexion hiérarchique rf pour les mpsoc," in *ComPAS'2014 : Conférence d'informatique en Parallélisme, Architecture et Système*, 2014, pp. track–architecture.

[7] U. S. Jha and R. Prasad, *OFDM towards fixed and mobile broadband wireless access*. Artech House, Inc., 2007.

[8] M. Hamieh, M. Ariaudo, S. Quintanel, and Y. Louët, "Sizing of the physical layer of a rf intra-chip communications," in *21st IEEE International Conference on Electronics Circuits & Systems*. IEEE, 2014, pp. 139–144.

[9] A. Brière, J. Denoulet, A. Pinna, B. Granado, F. Pêcheux, E. Unlu, Y. Louët, and C. Moy, "A dynamically reconfigurable rf noc for many-core," in *Proceedings of the 25th edition on Great Lakes Symposium on VLSI*. ACM, 2015, pp. 139–144.

[10] J. Palicot, C. Moy, M. Debbah, R. Couillet, H. Tembine, R. Séguier, D. Le Guennec, W. Jouini, G. Tourneur, Y. Louët *et al.*, "De la radio logicielle à la radio intelligente," 2010.

[11] K. Seong, R. Narasimhan, and J. M. Cioffi, "Queue proportional scheduling via geometric programming in fading broadcast channels," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1593–1602, 2006.

[12] Y. Pan, J. Kim, and G. Memik, "Tuning nanophotonic on-chip network designs for improving memory trafics," *PICA@ MICRO2009*, 2011.