Building analytical model for latency in Network-on-Chip using queuing theory

Dr. Radwan Dandah [*]
Dr. Talal Al-Aateky [**]
Rony Kassam [***]

(Received 30 / 1 / 2019. Accepted 13 / 5 / 2019)

□ ABSTRACT □

The Network-on-Chip architectures suffer from the difficulty of allocating processing resources in the execution time to suit the complex applications implemented on these architectures. In the event of any change in the input traffic, there is no change in the architecture level and its configuration. So it will be an inefficient running of the application on the architecture. Researches are currently modeling the input rates for data and determining service times required to implement application tasks [1]. These researches are limited to the NoC mechanisms of handling with different models but without any modifications to the architecture to suit these models. Therefore, this research propose modifications on the mathematical models used to include changes in the queue lengths and resource utilization rates to ensure the compatibility with application status. These status are classifying with multiple classifications close to the application tasks and can be included within the architecture in such a way as to ensure the effective operation of the class occasion. The research resulted in a reduction of about 10% of the latency for low input rates with the possibility of dealing more with the increasing of input rates, as well as opening up the application to reflect its effect in a way that can be improved on the architecture through the classification included in the mathematical model.

Keywords: Network-on-Chip, Latency mathematical model, Queuing theory, Application status, Classification.

[*] **Professor – Department of Computer Systems and Networks – Faculty of Information Engineering – Tishreen University – Lattakia – Syria**

[**] **Assistant Professor – Department of Computer Systems and Networks – Faculty of Information Engineering – Tishreen University – Lattakia – Syria**

[***] **PhD Student – Department of Computer Systems and Networks – Faculty of Information Engineering – Tishreen University – Lattakia – Syria**

بناء نموذج تحليلي للتأخير في الشبكة المضمّنة على الشريحة باستخدام نظرية الترتيل

الدكتور رضوان دنده [*]
الدكتور طلال العاتكي [**]
روني قسام [***]

(تاريخ الإيداع 30 / 1 / 2019. قُبِل للنشر في 13/ 5 / 2019)

□ ملخّص □

تعاني معماريات الشبكات المضمّنة على الشريحة من صعوبة توزيع موارد المعالجة في زمن التنفيذ بما يتناسب مع التطبيقات المعقدة المنفذة على هذه المعماريات، حيث أنه عند حدوث أي تغير في الدخل لا يؤدي إلى أي تغير على مستوى المعمارية والإعدادات الخاصة بها وإنما سيبقى التعامل حسب الإعدادات المثبتة سابقاً وبالتالي سيكون هناك تشغيل غير فعال للتطبيق على المعمارية. تقوم حالياً الدراسات بنمذجة معدلات الدخل للبيانات وتحديد أزمنة الخدمة المطلوبة لتنفيذ مهام التطبيق [1]، تقتصر هذه الدراسات على آليات تعامل الشبكات المضّمنة على الشريحة مع اختلاف هذه النماذج ولكن بدون أي تضمين لتعديلات ضمن المعمارية لتتناسب مع هذه النماذج. لذلك تم في هذا البحث اقتراح تعديلات على النماذج الرياضية المتبعة لتضمين تعديلات في أطوال الرتل ومعدلات استخدام الموارد وبشكل يؤمن توافق مع حالات التطبيق، حيث يتم تصنيف هذه الحالات بأصناف متعددة قريبة من مهام التطبيق من جهة وقابلة للتضمين ضمن المعمارية بشكل يؤمن التشغيل الفعال للصنف بأزمنة تأخير مناسبة. توصل البحث إلى تخفيض حوالي 10% من أزمنة التأخير بالنسبة لمعدلات دخل منخفضة مع إمكانية التعامل بشكل أكبر مع زيادة معدلات الدخل وكذلك فتح المجال للتطبيق ليعكس أثره بشكل يكون قابلاً للتحسين على المعمارية من خلال التصنيف المضمن بالنموذج الرياضي.

الكلمات المفتاحية: الشبكات المضمنة على الشريحة، نموذج رياضي للتأخير، نظرية الترتيل، حالات التطبيق، التصنيف.

[*] **أستاذ – قسم النظم والشبكات الحاسوبية – كلية الهندسة المعلوماتية – جامعة تشرين – اللاذقية – سورية.**
[**] **مدرس – قسم النظم والشبكات الحاسوبية – كلية الهندسة المعلوماتية – جامعة تشرين – اللاذقية – سورية.**
[***] **طالب دكتوراه – قسم النظم والشبكات الحاسوبية – كلية الهندسة المعلوماتية – جامعة تشرين – اللاذقية – سورية.**

مقدمة:

هناك العديد من التحديات في تصميم أنظمة الشبكات المضمّنة على الشريحة Network-on-Chips (NoC) المتعددة النوى، مثل تعقيد الـNoC بالإضافة إلى المتطلبات الصارمة لاستهلاك الطاقة والانتاجية. حيث إن تصميم نظام معتمد على الـNoC يتطلب العديد من خطوات التركيب Synthesis بما في ذلك توزيع المهام والتوجيه وتموضع النوى [2]. حيث أنه من أجل تمكين هذا النظام من التعامل مع حركة البيانات الخاصة بتطبيق معين، يحتاج المصمم أولاً إلى الجدولة وتعيين المهام على عناصر المعالجة المتاحة. بعدها يحتاج إلى اكتشاف تموضع النوى وكيفية حجز مسارات التوجيه. وهناك حاجة إلى أداة لتحليل الأداء من أجل تقييم فيما لو أن إعدادات الـNoC من أجل دخل تطبيق معين قادت إلى تصميم أفضل وذلك للمقارنة مع التصاميم الأخرى مع تلبية قيود التصميم في نفس الوقت. بتنفيذ محاكاة على شبكة معينة يمكن الحصول على نتائج لتقييم الأداء مع دقة عالية ومع ذلك فأنها تعاني من أوقات تقييم طويلة وهي مناسبة فقط لتقدير مجموعة فرعية صغيرة من المقاييس، وبسبب هذا يتم اعتماد نمادج أداء NoC على نطاق واسع. من بين جميع مقاييس أداء الـNoC يتم اعتبار التأخير Latency الأكثر أهمية لأنه يحدد انتاجية النظام بأكمله تحت أعباء معينة.

اقترح [1] نموذج لضبط أداء النظام من خلال معرفة معدل تأخير الدفقات في الـNoC المعتمدة على الأنظمة متعددة النوى والذي اعتمد على تشكيل ترتيل G/G/1/K حيث تم نمذجة وصول حركة البيانات الحالية باستخدام مقاربات بواسون، يتم توسيعها إلى توزيع أسي عام للوصول البيني للطرود والتي يمكن أن تمثل نمط لحركة بيانات دفقية. على غرار عملية خدمة الطرود داخل كل موجه NoC تكون النمذجة مع التوزيع العام لحساب ارتباط زمن الخدمة بين الموجهات وتدفقات حركة البيانات، كما استند [3] على تشكيلات نظرية الارتال G/G/1، ومساهمة كل قناة في التوجيه ضمن الـNoC. إن نماذج تقدير الأداء دقيقة مع الأخذ بعين الإعتبار افتراض أن طول الطرد يخضع للتوزيع الأسي، وبالتالي فإن وقت خدمة الطرد في جهاز التوجيه يوزع بشكل كبير، وأن زمن الوصول البيني لحركة البيانات يفترض أن يتبع توزيع بواسون. وقد لوحظ في [4] أن كثير من أنظمة الـNoC تتبع سلوك حركة البيانات نمط هندسي متكرر يعتمد على الشرائح طويلة المدى، و التوزيعات لزمن الخدمة متناغمة مع هذه الشرائح. وبالتالي فإن دقة النموذج القائم على نظرية الارتال يتعرض للخطر في هذه الحالات.

تم اقتراح نموذج تأخير M/G/1 [5] من أجل تحليل الـNoC. وهو يفترض على أن معدل وصول ترويسات الـflits (بدلاً من كامل الطرد) يتبع توزيع بواسون. تم تقديم نموذج تأخير G/G/1 [3] معتمد على NoC ذو أولوية ثابتة يقوم بنمذجة زمن وصول الدفقات باستخدام عملية بواسون المضمّنة بماركوف ثنائية الحالة 2-state Markov-Modulated Poisson Process (MMPP). مع ذلك، يستهدف هذا النهج بنية محددة للموجه على أساس الأولوية في حين أن العديد من موجهات الـNoC قد تستخدم التحكيم المنصف مثل Round Robin (RR). كذلك تم تقديم نموذج ترتيل M/G/1/N لموجهات Worm-Hole (WH) ضمن الـNoCs [6]، يفترض هذا النهج تدرج المخازن المؤقتة من حيث الطرود بدلاً عن الـflits و بالتالي يمكن لمخزن مؤقت واحد أن يحمل حتى N طرد خلال التحليل. قد لا يكون هذا الحل مناسباً بالنسبة للـNoCs بمخازن صغيرة نسبياً (فقط عدة flits) وذلك من أجل توفير المساحة والطاقة.

أهمية البحث وأهدافه:

تكمن أهمية البحث في عكس أثر التطبيق بشكل أكبر على الـNoC أثناء التنفيذ وكذلك على المستوى المعماري، وبالتالي إعطاء الـNoC قدرة التأقلم حتى على الطبقات العتادية مع تغييرات التطبيق على مدى الزمن واستخدام متغيرات التطبيق في أبحاث متخصصة بزيادة انتاجية التطبيق وبالتالي عكس أثره على انتاجية الـNoC بدون إعادة تطوير المعمارية وبالتالي تخفيض زمن الوصول إلى السوق Time to Market.

طرائق البحث ومواده:

تم استخدام نموذج عام لمعمارية الـNoC لاستيعاب الموجهات مع مخازن مؤقتة، مما يتيح النظر في تراكب طول الطرد مع حجم المخزن المؤقت. إن هذا الأطار المقترح هو عام تماماً ويمكن تطبيقه على أي مخطط شبكي للـNoC مع مختلف مجدولات المهام وخوارزميات التوجيه. ثم تم إدخال بارامترات جديدة إلى النموذج تتعلق بحالات التطبيق والأصناف التي تصف هذه الحالة والتي تغلف متغيرات الرتل في عقد الـNoC ومعدل استخدام الموارد، وبالتالي جعل هذه البارامترات قابلة للتعديل من طرف التطبيق والوصول إلى إنتاجية أفضل للـNoC والتي تتجلى في قيم منخفضة للتأخير. وأخيراً تم محاكاة هذا النموذج باستخدام المحاكي BookSim والقيام بالمقارنة مع نموذج مرجعي.

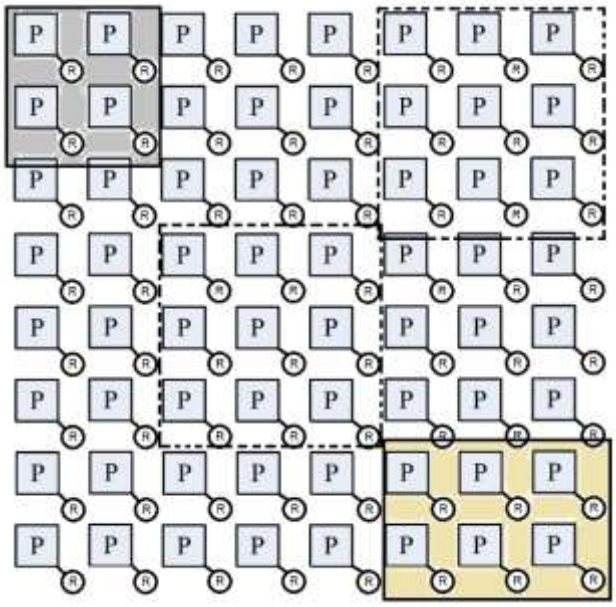
نموذج الـNoC التحليلي المقترح:

1- فرضيات:

تم افتراض أن تطبيقات الهدف تمت جدولتها ومطابقتها على منصة الـNoC الهدف وأن عنوان العقدة المصدر والهدف في التطبيق كلها معروفة. وأيضاً، تم استخدام فكرة نمذجة دفقات حركة البيانات في أنظمة الحواسيب المتعددة والمتصلة بشكل Mesh، حيث نفترض فيها بأن أزمنة الوصول البيني لطرود الدفق تم تمييزها باستخدام التوزيع الاسي العام GE، مع متوسط λ_f^{-1} و تشتت C_a^2 . لذلك فإن التمثيل $(\lambda_f^{-1}$ ، $C_a^2)$ لنموذج حركة البيانات يعتبر دخلاً للإطار التحليلي المقترح.

كما تم اعتماد معمارية الموجه Worm-Hole حيث يوجد مخزن مؤقت واحد على كل منفذ دخل. للبساطة نحن نفترض أن الطرود لها حجم ثابت m من الـflits، ومع ذلك فإن هذا الإفتراض يمكن أن يكون مريح لتغطية توزيع طول الطرد. الافتراضات الأخرى تتعلق أيضاً بمصادر حركة البيانات (عناصر المعالجة PEs المصدر) لديها حجم رتل لانهائي والوجهات تخدّم على الفور الـflits الواصلة.

كما نفترض بيئة محاكاة من أجل تقييم دقة وتحليل أداء التطبيق باستخدام قيم عددية، حيث تم وضع تصور للتطبيق بأنه يستطيع التعامل مع عدة حالات، كل دفق له حالته الخاصة، مثلاً كاميرا موجهة على مدخل مطار أو مجمع تجاري يتألف من عدة أرتال يدخلها الزبائن، ويتم تحليل ومعالجة المدخلات من كل رتل باستخدام مجموعة من عناصر المعالجة، كما في الشكل(1)، حيث يبين الشكل إسناد يدوي لمجموعة من عناصر المعالجة إلى رتل محدد، وبالتالي ستتم النمذجة على جعل الـNoC متوافقة أكثر ما يمكن من أجل هذا المخطط.



الشكل (1): ربط حالات التطبيق بموارد محددة من الـNoC

أي يجب أن تكون الحالة التي تتطلب أعلى تخديم ستأخذ أعلى سعة للرتل وبحسب عدد عناصر المعالجة المتاح في كل منطقة بشكل يؤدي إلى تخفيض التأخير باعتبار هذه الحالة من التطبيق أكثر أولوية في هذا المنطقة.

2- نمذجة التصنيف والأولوية:

بحسب هدف البحث، سيكون هناك حاجة إلى إدخال تصنيفات وأولويات للتطبيق الهدف، حيث يتم بدايةً تحديد هذه التصنيفات مع توزيع احتمالي، وتضمينها من خلال حالات الرتل G/G/1 [7].وبالتالي يصبح تصنيف التطبيق مقابل لحالة الرتل وبالتالي فإن مجموعة حالات الرتل ستحقق ما يلي:

$$\sum_{S\in Q} P(S) = 1 \qquad (1)$$

حيث Q مجموعة حالات الرتل، S حالة من حالات الرتل، وبالتالي يكون P(S) هو احتمال أن يكون الرتل G/G/1 في الحالة S. ونعرف أيضاً ρ_i معدل استخدام الحالة للموارد الرتل بالعلاقة:

$$\rho_i = \frac{\lambda_i}{\mu_i}$$

حيث i معدل وصول الطرود الوسطي للحالة، μ_i متوسط معدل الخدمة للحالة، ونعرف أيضاً C_a^2 التشتت في توزيع زمن الوصول البيني للحالة، C_s^2 التشتت في توزيع زمن الخدمة للحالة.

وبالتالي يصبح معدل استخدام الحالات للموارد تعطى بالعلاقة [7]:

$$\rho_i = \sum_{S\in Q} S_i(S)P(S) \quad \begin{cases} w = i \quad so \quad S_i(S) = 1 \\ w \neq i \quad so \quad S_i(S) = 0 \end{cases}$$

أي عندما تكون الرتل في الحالة i عندها ستكون القيمة لـ$S_i(S)$ تساوي 1 وإلا فهي 0. وكذلك يمكن الحصول على طول الرتل الوسطي بنفس الطريقة من خلال العلاقة:

$$QL_i = \sum_{S\in Q} QL_i(S)P(S) \quad QL_i(S) = QL_i \quad where \quad QL_i > 0$$

3- نمذجة حركة البيانات ذات التوزيع الأسي العام:

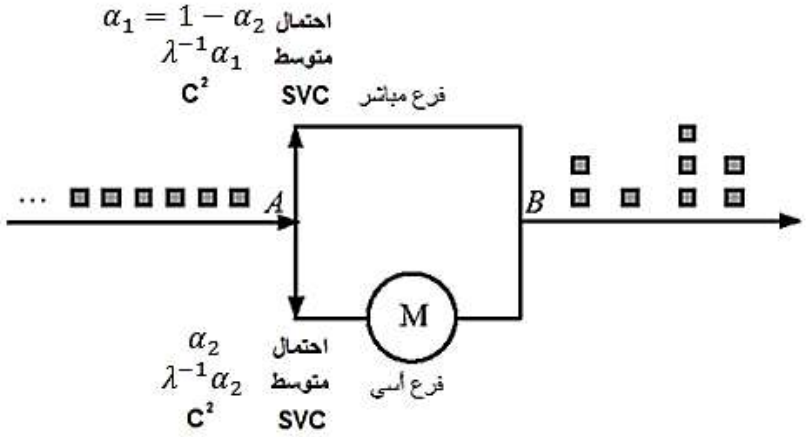
تظهر أنماط حركة البيانات في العديد من تطبيقات الـNoC على نطاق واسع من المقاييس الزمنية، حيث تعتبر عملية الترتيل فيها عملية عشوائية Stochastic (أو تدعى عملية ماركوف) كون خاصية عديم الذاكرة Memoryless تعود إلى خواص ماركوف، حيث يتميز رتل ماركوف بعملية وصول بواسون وأزمنة خدمة أسية تحقق تشتت عالي Maximum Entropy للأرتال العامة، لذلك يمكن الانتقال بسهولة من G/G/1 إلى GE/G/1 [7]. يتم استخدام التوزيع الأسي العام Generalized Exponential (GE) لتمثيل حركة البيانات عند الـPEs المصدرية والوصلات، وعلاوة على ذلك فقد ثبت أن توزيع GE يوفر أيضاً مقاربة فعالة لحركة البيانات قصيرة أو طويلة المدى في الحواسيب الفائقة [1]. فيما يلي نستعرض بإيجاز توزيع GE، تابع التوزيع التراكمي Cumulative Distribution Function (CDF) لزمن الوصول البيني للطرود X يعطى بالعلاقة:

$$F(t) = P(X \leq x) = 1 - \alpha_2 e^{-ht},\ t \geq 0 \qquad (2)$$

بحيث:

$$h = \alpha_2\lambda = \frac{2\lambda}{c^2+1} \qquad \alpha_1 = 1 - \alpha_2 = \frac{c^2-1}{c^2+1} \qquad \alpha_2 = \frac{2}{c^2+1}$$

بمتوسط λ^{-1} وتشتت C^2للمتحول X. إن عملية توليد طرد GE مبينة بالشكل (2). حيث إن الطرد يأخذ زمن خدمة صفري للوصول إلى نقطة الإنطلاق (أي نقطة توليد الطرد B) مع احتمالية α_1.



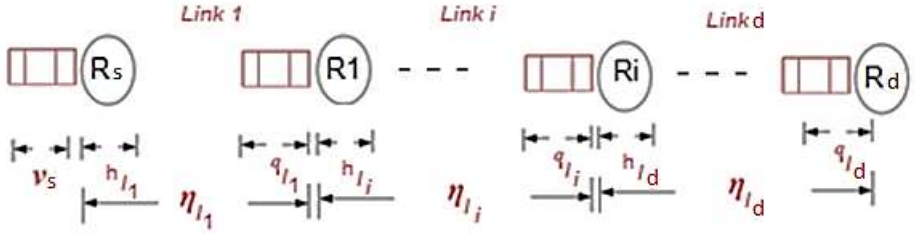
الشكل (2): توزيع أسي عام لحركة البيانات

أما بقية الإحتمال والمقصود هنا α_2 ، يحتاج الطرد لاجتياز النظام مع وقت خدمة موزعة بشكل أسي بمتوسط يساوي $\lambda^{-1}\alpha_2$. تتكون دفقة الطرود من طرد يأتي من فرع أسي (فرع M) بالإضافة إلى عدد من الطرود المتتابعة القادمة من فرع مباشر.

4- نمذجة الموجه:

بحسب [1] إن العناصر الرئيسية الثلاثة للنماذج التحليلية لـWH هي زمن نقل الـflit η ، و زمن اكتساب المسار h، و زمن ترتيل المصدر v_s، وسيتم الآن أخذ كل عنصر والتفصيل في آلية تقدير قيمته، وبعدها سيتم تحديد الإضافات على كل عنصر لتعديل النموذج التحليلي حسب هدف البحث:

* زمن نقل الflit: يعرّف زمن نقل الflit η للوصلة l_{ab} بأنه الزمن الذي تستغرقه الflit الرأس لمغادرة المخزن المؤقت في العقدة المصدر، والوصول إلى مقدمة المخزن المؤقت في الوصلة الحالية l_{ab} بعد حصوله على سماحية الوصلة. يوضح الشكل(3) مفهوم هذا الزمن.



الشكل (3): رسم توضيحي للأزمنة ضمن الـNoC

وبشكل أكثر تحديداً، يتكون من جزئين، الأول هو الزمن المناسب لترك الموجه المصدر. من أجل توجيه WH، يعتبر هذا الزمن قيمة ثابتة تعتمد على عدد مراحل الأنبوبية H_S في الموجه والوصلة. أما الجزء الثاني يحسب الزمن الذي تأخذه الflit الرأس للوصول إلى الجزء الأمامي من المخزن المؤقت للوصلة l_i (q_{l_l})، وموضح أيضاً بالشكل (3). يمكن تقريب قيمة هذا الزمن من خلال زمن الانتظار في نظام الترتيل GE/G/1. حيث متوسط معدل وصول الflit في نظام الترتيل هذا يعطى بالعلاقة:

$$\lambda_{flit}^{l_{ab}} = m \times \lambda_{packet}^{l_{ab}}$$

$$\lambda_{packet}^{l_{ab}} = \sum_{f \in F_{l_{ab}}} \lambda_f = \sum_{\forall s} \sum_{\forall d} (\lambda_{sd} \times R(s,d,l_{ab}))$$

حيث $R(s,d,l_{ab})$ هو قيمة ثنائية تشير إلى أن القناة هي عنصر من مجموعة مسار التوجيه $P_{s,d}$:

$$R(s,d,l_{ab}) = \begin{cases} 1 & if\ l_{ab} \in P_{s,d} \\ 0 & otherwise \end{cases}$$

يمكن حساب متوسط زمن خدمة الflit $(s_{flit}^{l_{ab}})$ في نظام الترتيل هذا من خلال معدل موزون لزمن الخدمة لجميع الدفقات التي تمر عبر الوصلة l_{ab} وتعطى بالعلاقة:

$$s_{flit}^{l_{ab}} = \frac{\sum_{\forall f \in F_{l_{ab}}} \left[\lambda_f \times \left(\frac{h_{l_{i+1}^f}}{m} + 1 \right) \right]}{\sum_{\forall f \in F_{l_{ab}}} \lambda_f} \qquad (3)$$

تبين العلاقة (3) أنه من أجل طرد مكون من m flits، والتي تأخذ $h_{l_{i+1}^f}$ دورة من أجل الflit الرأس في المخزن المؤقت للفوز بالوصلة التالية في الدفق f. بعد ذلك نحتاج إلى $1 + h_{l_{i+1}^f}$ دورة لخدمة الflit الرأس. بعد ذلك، يتم نقل جسم وذيل الflits في دورة واحدة بدون أي تأخير إضافي. وبالتالي يتم حساب متوسط زمن خدمة الflit للطرد بأكمله كما يلي $1 + \frac{h_{l_{i+1}^f}}{m}$. بعد الحصول على متوسط معدل الوصول $\lambda_{flit}^{l_{ab}}$ ومتوسط زمن الخدمة $s_{flit}^{l_{ab}}$ في الرتل، يمكن تطبيق صيغة الترتيل GE/G/1 لمقاربة q_{l_l}. ويمكن الحصول على $\eta_{l_i^f}$ من خلال العلاقة $\eta_{l_i^f} = q_{l_l} + H_s$. أي اعتمد [1] على التابع $\mathrm{Y}(s_{flit}^{l_{ab}}, \lambda_{flit}^{l_{ab}})$ لاستخلاص زمن q_{l_l} ، بينما تم الاعتماد في البحث على $\overline{\mathrm{Y}}(\lambda_{flit}^{l_{ab}}, QL_k)$ ، حيث QL_k هو طول الرتل الوسطي للحالة k.

* زمن اكتساب المسار: إن زمن اكتساب المسار h للدفق f في القفزة i للوصلة l_{ab} (أي l_i^f)، يعرف بأنه زمن انتظار الـFlit الرأس لمنح سماحية الدخول إلى المخازن المؤقتة في l_{ab} بعد التنافس مع الدفقات الأخرى الموجهة نحو نفس الخرج، وعادة ما يتم النمذجة كزمن الانتظار في نظام الترتيل مثل G/G/1، GE/G/1 [1]. إذا لم تزدحم الوصلات المصدر، فإن زمن الخدمة في هذه الحالة يساوي طول الطرد (أي m دورة) وذلك لأن الطرد بأكمله يمكن أن يعبر بطريقة مستمرة. ومع ذلك، عندما يكون هناك انسداد شديد على طول الطريق، يحدث السيناريو الأسوأ عندما تصل ترويسة الطرد إلى الوصلة l_d في الشكل (3) و تكون الأماكن من المخازن المؤقتة المتراكمة من مقدمة المخزن المؤقت في الوصلة l_1 إلى نهاية المخزن المؤقت في الوصلة l_d تكفي فقط لكامل الطرد. لذلك قام [1] بربط زمن الخدمة $s_{l_i^f}$ بزمن تأخير الطرد من نهاية المخزن المؤقت في الوصلة l_s إلى نهاية المخزن المؤقت في الوصلة l_d. حيث عندما لا تكون قنوات المصدر مزدحمة على طول مسار التوجيه، يقترب زمن الخدمة l_i^f من طول الطرد m. وبشكل آخر في حال وجود إعاقة كبيرة في القفزات اللاحقة، تقترب $s_{l_i^f}$ من زمن تأخير الطرد الناتج عن تأخير الإزدحام في الوصلات التالية.

استنادا إلى المناقشات المذكورة أعلاه حيث يتم الحصول على زمن اكتساب المسار من زمن خدمة الوصلة، أي اعتمد [1] على التابع Z($s_{l_i^f}$) لاستخلاص زمن اكتساب المسار، بينما تم الاعتماد في البحث على $\bar{Z}(\rho_k, \lambda_{flit}^{l_{ab}})$ ، حيث ρ_k هي معدل استخدام الحالة k للرتل.

* زمن ترتيل المصدر:

بحسب [1] تم نمذجة رتل المصدر في واجهة الشبكة NI كنظام الترتيل مع سعة لا نهائية ولكن نحتاج إلى تقليل السعة وفقاً لمتطلبات التطبيق وبالتالي تم التعامل مع v_s المتعلق بزمن الخدمة ومعدل الوصول، تم في البحث التعامل مع تأخير الترتيل W_k للحالة k المتعلقة بطول الرتل المصدري QL_k وكذلك بمعدل استخدام الرتل ρ_k ومعدل وصول البيانات.

5- تشكيل علاقة التأخير نهاية لنهاية:

في الـNoC ذات الموجهات WH، يعبر عن تأخير الدفق نهاية لنهاية $L_{s,d}$ من أجل دفق معين $f_{s,d}$ ، بالعلاقة [1] [8]:

$$L_{s,d} = v_s + \eta_{s,d} + h_{s,d} \quad (4)$$

من أجل حساب $h_{s,d}$، نحتاج إلى اعتبار زمن إكتساب المسار لكل وصلة l_i^f تقع على مسار التوجيه للدفق f ، لذلك: $h_{s,d} = \sum_{i=1}^{d_f} h_{l_i^f}$

حيث $h_{l_i^f}$ هو زمن منافسة ترويسة الطرد للحصول على قناة مع تدفقات أخرى لوصلة l_i^f وبطول مسار توجيه d_f (عدد القفزات) للدفق f. إذا أشرنا للزمن اللازم لإرسال الـflit الرأس بـ $\eta_{l_i^f}$ ، عندها يمكن إعادة كتابة زمن نقل الطرد كما يلي:

$$\eta_{s,d} = \sum_{i=1}^{d_f} \eta_{l_i^f} + (m-1) \quad (5)$$

حيث الجزء الأول يدل على زمن انتقال الـflit الرأس عبر الشبكة، أما الجزء الثاني يمثل بشكل تقريبي زمن تسلسل الـflits الجسم والذيل للطرد [1] [8].

من أجل استخلاص العناصر $L_{s,d}$ ، $h_{l_i^f}$ و $\eta_{l_i^f}$ لكل وصلة والتي هي عناصر مسار التوجيه $f_{s,d}$، يجب تطوير النموذج المفصل للحصول على h و η لكل وصلة من أجل تحديد الأزمنة بشكل أكثر دقة، حيث بالنسبة لـ η يمكن التعبير عنها من خلال q التي تعبر عن تأخير الـFlit للوصول إلى المخزن المؤقت الخاص بالدخل بالموجه التالي، وكذلك من خلال Hs الذي يعبر عن زمن خدمة الـFlit مع النقل، لذلك تم أخذ شبكة NoC كما في الشكل (3) وتحليل التأخير فيها. بداية مع العلاقة (4) نكتب [1] [8]:

$$L_{s,d} = v_s + \sum_{i=1}^{d_f} \eta_{l_i^f} + (m-1) + \sum_{i=1}^{d_f} h_{l_i^f}$$

ولكن باعتبار أن زمن نقل الطرد η يضم كل من H_s زمن خدمة الـFlit مع النقل وزمن q تأخير وصول الـFlit إلى مقدمة المخزن المؤقت الخاص بدخل الموجه التالي، لذلك نكتب:

$$L_{s,d} = v_s + \sum_{i=1}^{d_f} [\, q_{l_i^f} + H_s + (m-1)] + \sum_{i=1}^{d_f} h_{l_i^f}$$

$$L_{s,d} = v_s + \sum_{i=1}^{d_f} [\, Y(s_{l_i^f}, \lambda_{s_{l_i^f}}) + H_s + (m-1)] + \sum_{i=1}^{d_f} \mathrm{Z}(s_{l_i^f}) \qquad (6)$$

بحسب الشكل (3) وبعد قولبة الفكرة حسب الشكل (4) وإعتمادية القيم وبفرض لدينا k حالة ضمن التطبيق. عندها بتركيب معدلات استخدام الحالات للموارد مع بعضها كمركب ρ :

$$\rho = (\rho_1 \ \rho_2 \ \rho_3 \ \rho_4)$$

وبالأخذ بعين الاعتبار نموذج توزيع GE حسب العلاقة (2) ، نعرف y بأنها متحول توزيع عشوائي (0,1):

$$\mathrm{t} = F^{-1}(y) = \frac{1}{h}\ln\left(\frac{\alpha_1}{1-y}\right) \qquad (7)$$

وبالتعامل مع طول الرتل على أنه طول البيانات التي تملأ شبكة الـNoC بحسب عناصر المعالجة المتاحة [9]، وكذلك بعد تحليل طرق الترتيل وكيفية التنسيق بين أطوال الأرتال المختلفة التي تعمل ضمن نفس النظام وبالتنسيق بينها [10]، تصبح العلاقات التي تحدد أطوال الرتل الوسطية كما يلي:

$$\mathrm{QL}_1 = \rho_1 + \frac{\rho_1}{2(1-\rho_1)}\left(C_{\alpha_1}^2 - 1\right) + \frac{1}{2(1-\rho_1)}\left(\alpha_1 + \beta_1\right) \qquad (8)$$

$$\alpha_1 = {\rho_1}^2\left(C_{s_1}^2 + 1\right) \qquad \beta_1$$

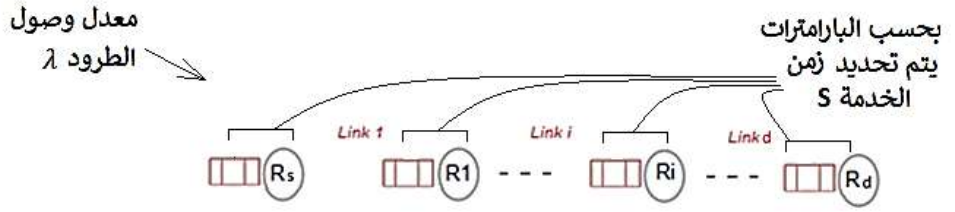
$$\mathrm{QL}_2 = \rho_2 + \frac{\rho_2}{2(1-\rho_1-\rho_2)}\left(C_{\alpha_2}^2 - 1\right) + \frac{1}{2(1-\rho_1)(1-\rho_1-\rho_2)}\left(\alpha_2 + \beta_2\right) \qquad (9)$$

$$\alpha_2 = {\rho_2}^2\left(C_{s_2}^2 + 1\right) + \frac{\lambda_2}{\lambda_1}{\rho_1}^2\left(C_{s_1}^2 + C_{\alpha_1}^2\right) \qquad \beta_2 = \frac{\lambda_2}{\lambda_3}{\rho_3}^2\left(C_{s_3}^2 + 1\right) + \frac{\lambda_2}{\lambda_4}{\rho_4}^2\left(C_{s_4}^2 + 1\right)$$

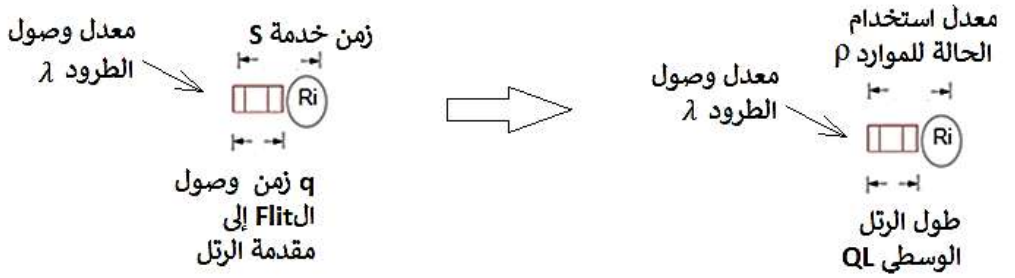
$$\mathrm{QL}_3 = \rho_3 + \frac{\rho_3}{2(1-\rho_1-\rho_2-\rho_3)}\left(C_{\alpha_3}^2 - 1\right) + \frac{1}{2(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)}\left(\alpha_3 + \beta_3\right) \qquad (10)$$

$$\alpha_3 = \rho_3{}^2\left(C_{s_3}^2+1\right)+\frac{\lambda_3}{\lambda_1}\rho_1{}^2\left(C_{s_1}^2+C_{\alpha_1}^2\right)+\frac{\lambda_3}{\lambda_2}\rho_2{}^2\left(C_{s_2}^2+C_{\alpha_2}^2\right) \qquad \beta_3=\frac{\lambda_3}{\lambda_4}\rho_4{}^2\left(C_{s_4}^2+1\right)$$

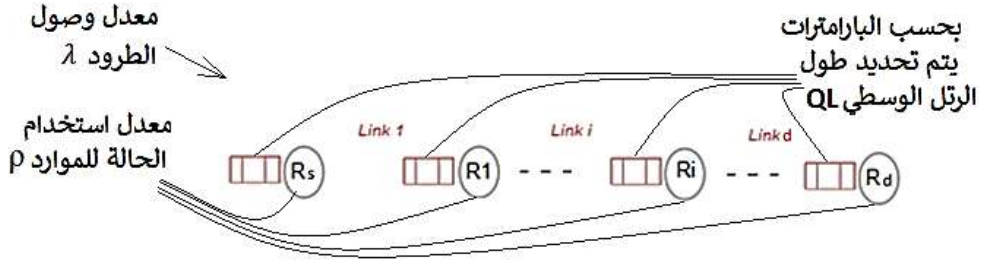
حسب النموذج المرجعي [1]



المطابقة المستخدمة في البحث



النموذج المقترح



الشكل (4): وصف مبسط لفكرة النموذج المقترح

$$QL_4=\rho_4+\frac{\rho_4}{2(1-\rho_1-\rho_2-\rho_3-\rho_4)}\left(C_{\alpha_4}^2-1\right)+\frac{1}{2(1-\rho_1-\rho_2-\rho_3)(1-\rho_1-\rho_2-\rho_3-\rho_4)}\left(\alpha_4+\beta_4\right) \qquad (11)$$

$$\alpha_4 = \rho_4{}^2\left(C_{s_4}^2+1\right)+\frac{\lambda_4}{\lambda_1}\rho_1{}^2\left(C_{s_1}^2+C_{\alpha_1}^2\right)+\frac{\lambda_4}{\lambda_2}\rho_2{}^2\left(C_{s_2}^2+C_{\alpha_2}^2\right)+\frac{\lambda_4}{\lambda_3}\rho_3{}^2\left(C_{s_3}^2+C_{\alpha_3}^2\right) \qquad \beta_4=0$$

وبالتالي بتعميم العلاقة على K حالة يصبح النموذج المتبع لحساب طول الرتل الوسطي المتأثر ببقية الحالات السابقة:

$$QL_k = \rho_k + \frac{\rho_k}{2(1-\rho_1-\rho_2-\rho_3-\rho_4-\cdots-\rho_k)}\left(C_{\alpha_k}^2 - 1\right) + \frac{1}{2(1-\rho_1-\rho_2-\rho_3-\cdots-\rho_{k-1})(1-\rho_1-\rho_2-\rho_3-\rho_4-\cdots-\rho_k)}\left(\alpha_k + \beta_k\right) \quad (12)$$

$$\alpha_k = {\rho_k}^2\left(C_{s_k}^2 + 1\right) + \frac{\lambda_k}{\lambda_1}{\rho_1}^2\left(C_{s_1}^2 + C_{\alpha_1}^2\right) + \frac{\lambda_k}{\lambda_2}{\rho_2}^2\left(C_{s_2}^2 + C_{\alpha_2}^2\right) + \frac{\lambda_k}{\lambda_3}{\rho_3}^2\left(C_{s_3}^2 + C_{\alpha_3}^2\right) + \cdots + \frac{\lambda_k}{\lambda_{k-1}}{\rho_{k-1}}^2\left(C_{s_{k-1}}^2 + C_{\alpha_{k-1}}^2\right) \quad \text{if} \quad k>1$$

$$\beta_{k-j} = \begin{cases} \frac{\lambda_{k-j}}{\lambda_{k-j+1}}{\rho_{k-j+1}}^2\left(C_{s_{k-j+1}}^2 + 1\right) + \cdots + \frac{\lambda_{k-j}}{\lambda_k}{\rho_k}^2\left(C_{s_k}^2 + 1\right) & for\ 1 < k-j < k \\ null \quad for \quad k-j = 1 \\ 0 \quad for \quad k-j = k \end{cases}$$

ويمكن حساب تأخير الترتيل في كل حالة كما يلي:

$$W_k = \frac{QL_k}{\lambda_k}$$

من أجل الحالة K ولكل وصلة i، وبعد مقابلة التابع Y بالتابع $\bar{}$ والتابع Z بالتابع $\bar{Z}$ [9] وكذلك حسب الشكل (4) تكون علاقة النموذج المطورة:

$$L_{s,d} = W_K + \sum_{i=1}^{d_f}[\bar{Y}(\lambda_{flit}^{l_{ab}}, QL_k) + H_s + (m-1)] + \sum_{i=1}^{d_f}\bar{Z}(\rho_k, \lambda_{flit}^{l_{ab}})$$

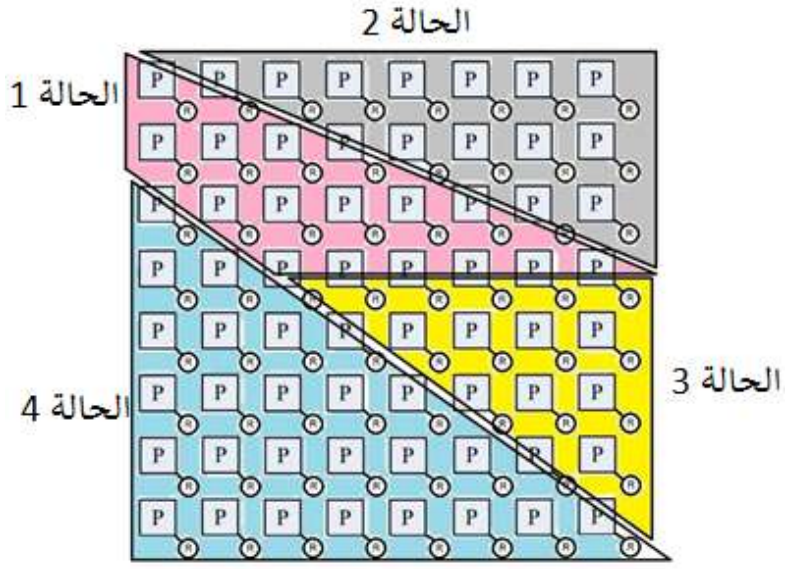
$$L_{s,d} = W_K + \sum_{i=1}^{d_f}[\frac{QL_{ik}}{\lambda} + H_s + (m-1)] + \sum_{i=1}^{d_f} m \times (\frac{\lambda}{\square_{ik}} - 1)$$

وبالتالي حسب العلاقة، يمكن حساب التأخير نهاية إلى نهاية لكل صنف وبشكل يؤثر كل صنف على الآخر حسب التوزيع العشوائي GE الذي يعطي قيم معينة لشمولية كل حالة للموارد المطلوبة، حيث إن قيم QL و W تختلف باختلاف تموضع الحالات ضمن شبكة الـNoC.

النتائج والمناقشة:

للتأكد من دقة النموذج يجب تمثيله بقيم عددية ضمن المحاكي، حيث تم استخدام BookSim 2.0 لمحاكاة شبكة الـNoC من النوع Mesh 8x8 أي تكون عقد التوجيه مربوطة كما الشكل(1)، تم وضع المخطط المرجعي حسب [1] وذلك بإعادة استخدام نفس قيم بارامترات الدخل ونمط حركة البيانات العشوائي Random مع معدلات حقن الطرود بواسون في كل عقدة مصدر. ويفترض أن طول الحزمة m لتكون ثابتة من 4 flits.

تم تحديد عدد حالات التطبيق بـ4 ومجال الحالة حسب الشكل(5)، لأنه من خلال التجربة تبين عند استخدام حالة أو حالتين فإنه لن يتم عكس أثر على التجربة بشكل واضح وذلك حسب العلاقات (8) (9) (10) (11) (12) ، لذلك عند اختيار أربع أصناف أو أكثر سيكون التأثير على النتيجة أفضل.



الشكل (5): توزع الحالات ضمن الـNoC في التجربة

يفترض النموذج أن يكون معدل وصول الطرود أصغر من معدل التخديم وتم ضبط ذلك من خلال معدل استخدام الحالات للموارد ρ حيث يوضح الجدول (1) هذه القيم والتي تم اعتمادها حسب [1] [11] وبشكل تكون الزيادة تدريجية وبحسب العلاقة(2). كما يفترض أن الحالة التي تتطلب أعلى تخديم سيأخذ أعلى سعة للرتل وبحسب عدد عناصر المعالجة المتاح.

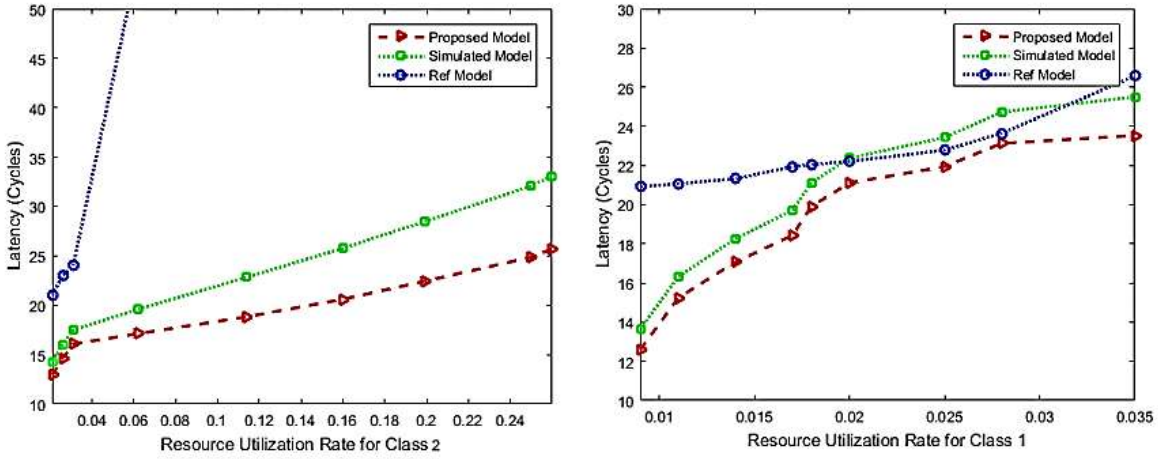
الجدول (1): معدل استخدام الحالات للموارد بشكل مركب وكل حالة على حدى

ρ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95
ρ_1	0.009	0.011	0.014	0.017	0.018	0.02	0.025	0.028	0.035
ρ_2	0.021	0.026	0.031	0.062	0.114	0.16	0.199	0.25	0.26
ρ_3	0.042	0.054	0.1	0.145	0.168	0.2	0.236	0.252	0.27
ρ_4	0.128	0.209	0.255	0.276	0.3	0.32	0.34	0.37	0.385

يبين الشكل (6 – a) الحالة الأولى من التطبيق كما في الشكل(5)، وباستخدام المعدل ρ_1 من الجدول(1)، نجد أنه مع معدلات منخفضة لاستخدام الحالة للموارد، سيبدي النموذج المقترح تفوق في التأخير على النموذج المرجعي وذلك بسبب كون النموذج المقترح لا يتعامل مع كامل شبكة الـNoC وإنما مع عدد محدد حسب الشكل(5)، ولكن مع زيادة المعدل سيتقارب التأخير وسبب ذلك عدم قدرة العدد المحدد من عناصر المعالجة على تخديم البيانات على الرغم من قيامة بزيادة طول الرتل الوسطي بشكل يحقق توزيع هذه البيانات بشكل أكبر على أرتال عناصر المعالجة.

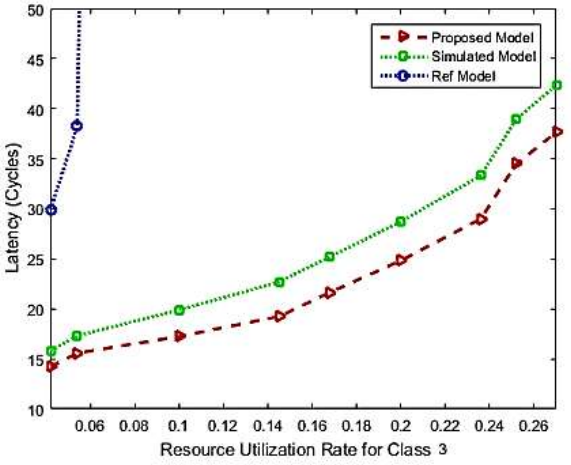
أيضاً يبين الشكل (6 – b و c و d) أنه مع زيادة معدل استخدام الحالات للموارد، فإن النموذج المرجعي يصل إلى عتبة معينة وبعدها يبدأ التأخير بالزيادة بشكل كبير، بينما النموذج المقترح يبين تحمل أكبر لهذه الزيادة وخاصة كون

عدد عناصر المعالجة وأطوال الأرتال الوسطية تساعد بشكل كبير على تحسين قيمة التأخير ولكن أيضاً تصل إلى عتبة معينة بعدها يبدأ التأخير بالزيادة بشكل كبير.

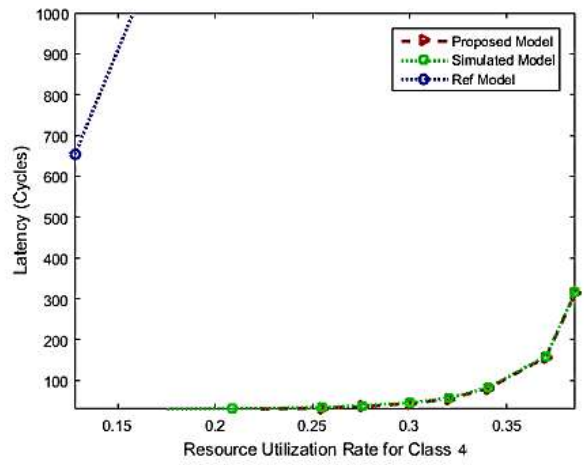


a – التأخير في الحالة الأولى

b – التأخير في الحالة الثانية



c – التأخير في الحالة الثالثة



d – التأخير في الحالة الرابعة

الشكل (6): مقارنة التأخير بين النموذج المقترح والمحاكاة والنموذج المرجعي

الاستنتاجات والتوصيات:

تم تطوير نموذج رياضي يمكنه التعامل مع حركة البيانات الواردة إلى شبكة الNoC بشكل فعال بحيث يتم إسناد هذه البيانات إلى عدد معين من عناصر المعالجة وبأطوال أرتال تحقق ما يناسب كمية البيانات المراد تخديمها، وبالتالي عند القيم الدنيا لحركة البيانات سيكون هناك انخفاض ما يقارب 10% في التأخير، وكذلك عند القيم العليا لكمية البيانات سيكون هناك تحمل أكبر والمحافظة على عمل الشبكة. من جهة أخرى أصبح هناك إمكانية لعكس أثر

التطبيق بشكل أكبر على الـNoC أثناء التنفيذ وعلى المستوى المعماري. حيث أصبح هناك مجموعة من البارامترات التي تؤثر في التأخير مباشرة ونابعة من البنية المعمارية لشبكة الـNoC.

المراجع:

[1] Z. QIAN; D. JUAN; P. BOGDAN; C. TSUI; D. MARCULESCU; R. MARCULESCU, *A Support Vector Regression (SVR)-Based Latency Model for Network-on-Chip (NoC) Architectures.* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 35, no. 3, 2016, 471-484

[2] R. KASSAM; T. ALAATEKY; R. DANDAH, *Enhancing virtual switching system of on-chip networks.* International Journal of Scientific & Engineering Research, vol. 6, no. 9, 2015,1888-1894.

[3] A. E. KIASARI; Z. LU; A. JANTSCH, *An analytical latency model for networks-on-chip.* Very Large Scale Integration VLSI Systems, IEEE Trans, vol.PP, no. 99, 2012, 1-11.

[4] V. D. NGO; J. Y. CHANG; Y. BAE; H. CHO; H. CHOI, *Latency Optimization for NoC Design of H.264 Decoder Based on Self-similar Traffic Modeling.* Parallel and Distributed Processing and Applications. ISPA, vol. 4742, 2007, 289-302.

[5] U. OGRAS; P. BOGDAN; R. MARCULESCU, *An analytical approach for network-on-chip performance analysis*. Computer-Aided Design of Integrated Circuits and Systems, IEEE Trans, vol. 29, no. 12, 2010, 2001-2013.

[6] M. LAI; G. LEI; X. NONG; W. ZHIYING, *An accurate and efficient performance analysis approach based on queuing model for network on chip.* International Conference on Computer-Aided Design ACM, 2009, 563-570.

[7] A. E. KIASARI; D. RAHMATI; H. SARBAZI-AZAD; S. HESSAB, *A Markovian performance model for networks-on-chip.* In Parallel, Distributed and Network-Based Processing, PDP, 16th Euromicro Conference on, 2008, 157-164.

[8] Z. QIAN, *High Performance Network-on-Chips (NoCs) Design: Performance Modeling, Routing Algorithm and Architecture Optimization*. arXiv, 2014, 1406-3790.

[9] J. FAHIMEH; J. AXEL; L. ZHONGHAI, *Weighted Round Robin Configuration for Worst-Case Delay Optimization in Network-on-Chip.* IEEE Transactions on Very Large Scale Integration VLSI Systems, vol. 24, no. 12, 2016, 3387-3400.

[10] HILLSTON, *Solving Queueing Models*, 2009, 10 Jun. 2018. *www.inf.ed.ac.uk/teaching/courses/ms/notes/solvingqueues.pdf*

[11] D. GHASSAN, M. NASEEF, S. SAJEED and G. AMLAN, *PaSE: A parallel speedup estimation framework for Network-on-Chip based multicore systems*. In Green and Sustainable Computing Conference IGSC 2017 Eighth International IEEE, 2017, 1-6.