RB-OLITS: A Worst-case Reorder Buffer Size Reduction Approach for 3D-NoC

Zhenmin Li, Yuqing Ma, Gaoming Du, Xiaolei Wang, Yukun Song, Duoli Zhang

School of Microelectronics, Hefei University of Technology, Hefei, China

Outline

- **Background**
- Preliminaries
- Worst-case Size Reduction of Reorder Buffer
- Experiments and Results
- **Conclusion**

Background

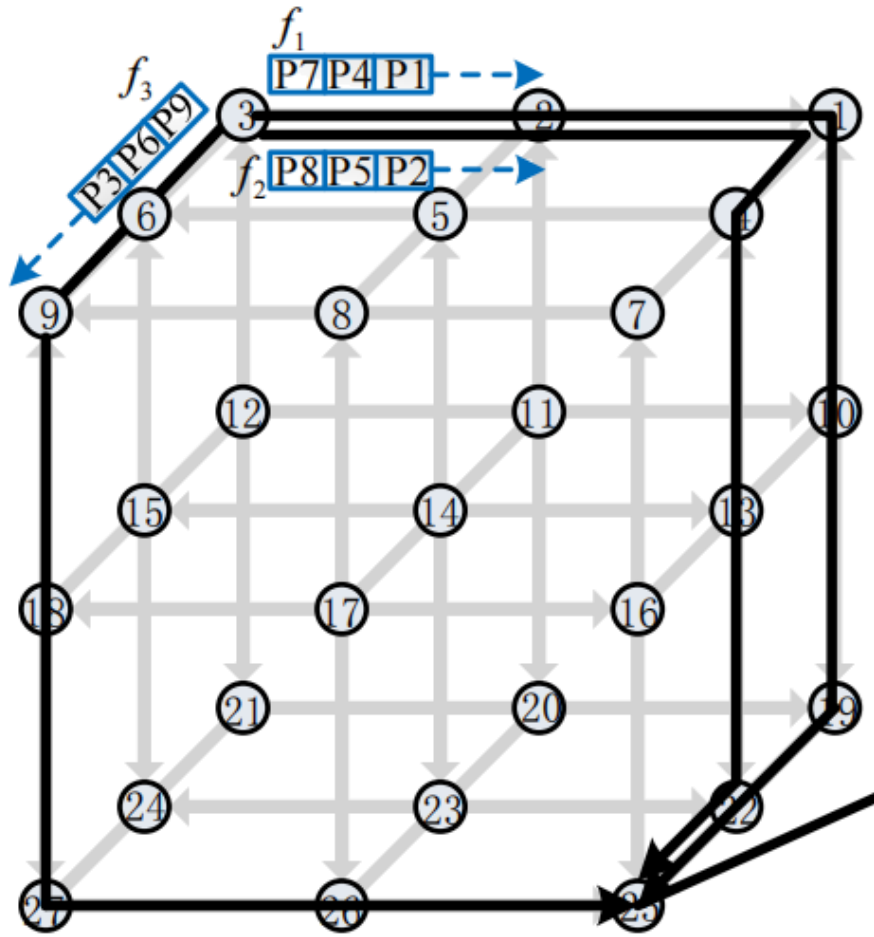
- 3D-NoC is emerging as a promising successor to the traditional bus-interconnecting architecture in MPSoC designs
- The packet based 3D-NoC paradigm is highly scalable and fundamentally decouples communication from computation
- In-order packet delivery is crucial for a majority of applications, for instance, multimedia or cache coherence protocols
- Packets arriving at the destination may cause the out-of-order problem
- Calculation and reduction of worst-case reorder buffer size has become a pivotal research topic.

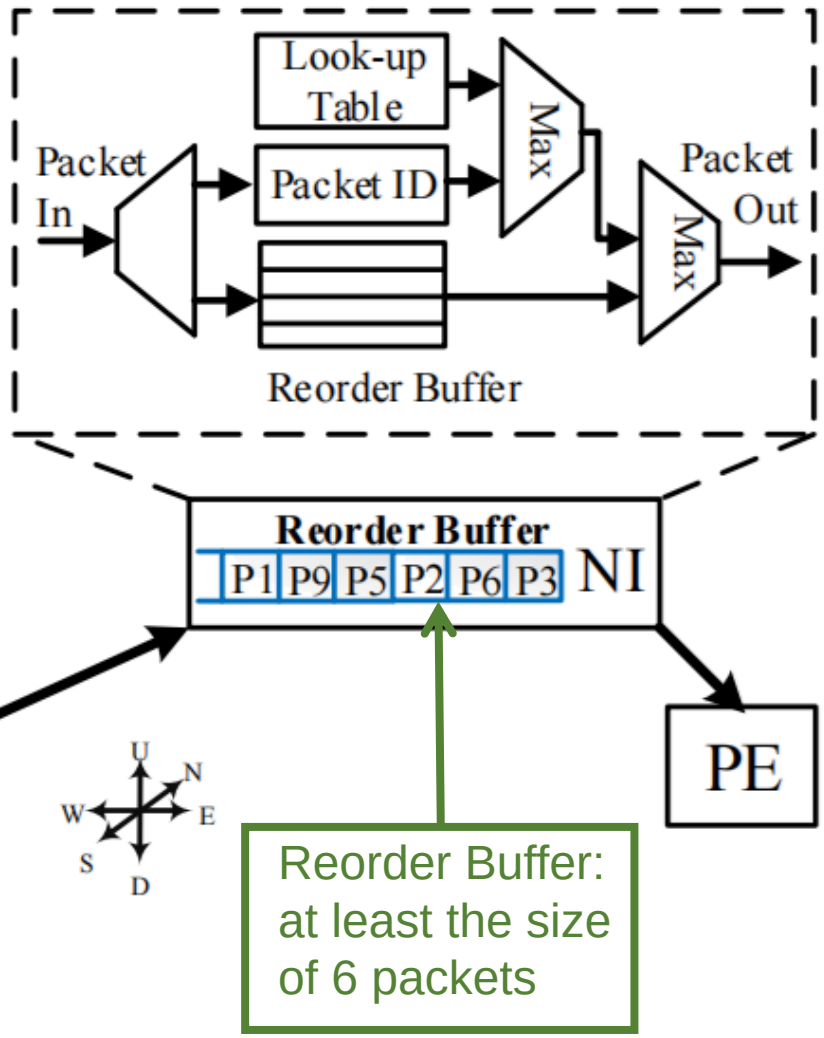
Outline

- Background
- **Preliminaries**
- Worst-case Size Reduction of Reorder Buffer
- Experiments and Results
- Conclusion

Preliminaries

sub-flows:
f1, f2 , f3



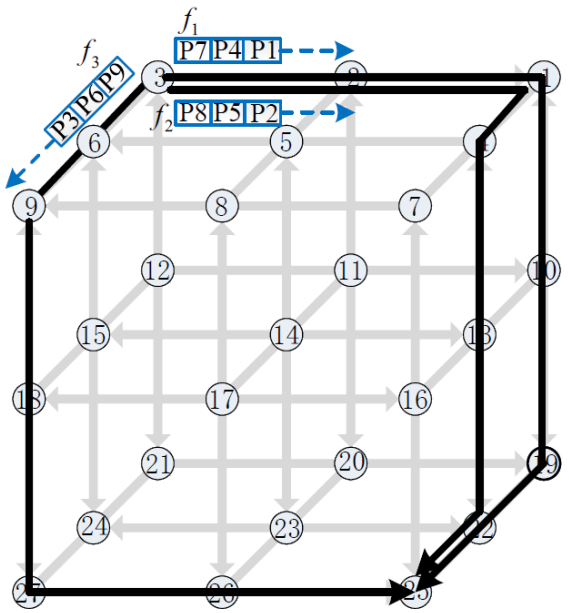


Outline

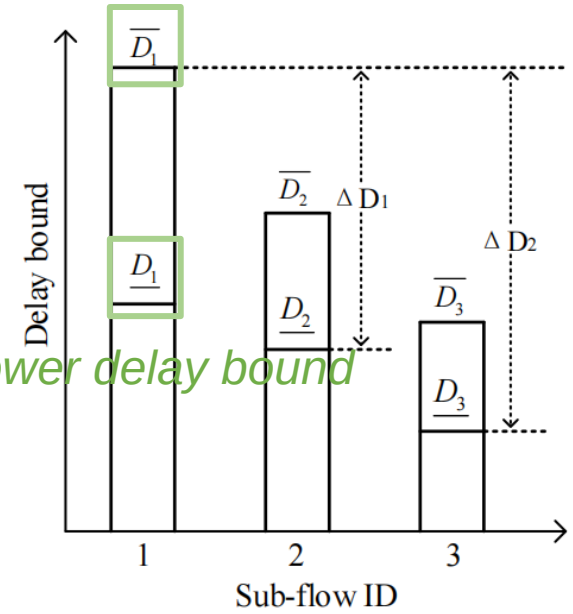
- Background
- Preliminaries
- **Worst-case Size Reduction of Reorder Buffer**
- Experiments and Results
- Conclusion

Worst-case Size Reduction of Reorder Buffer

Motivating Example



upper delay bound



lower delay bound

$$S_{rb} = \frac{(\overline{D}_1 - \underline{D}_2)}{\Delta t_2} + \frac{(\overline{D}_1 - \underline{D}_3)}{\Delta t_3}$$

generalize to n

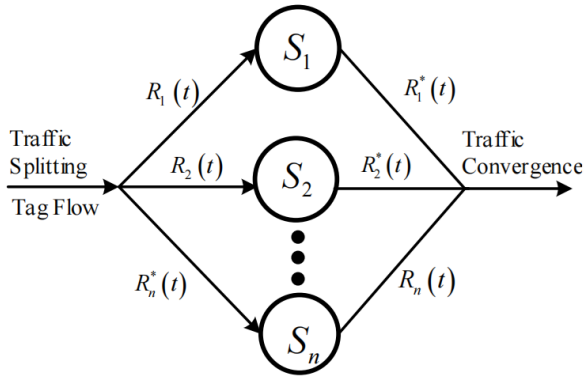
$$S_{rb}^{\max} = \frac{(\bar{D}_{\max} - \underline{D}_1)}{\Delta t_1} + \frac{(\bar{D}_{\max} - \underline{D}_2)}{\Delta t_2} + \cdots + \frac{(\bar{D}_{\max} - \underline{D}_n)}{\Delta t_n}$$

$$\bar{D}_{\max} = \max\{\overline{D}_1, \overline{D}_2, \cdots, \overline{D}_n\}$$

the packet injection intervals

Worst-case Size Reduction of Reorder Buffer

Worst-case Reorder Buffer Size Analysis Based on Network Calculus



Step 1: Traffic splitting

Step 2: Equivalent service curve (ESC) and delay bound calculation

Step 3: Calculation of worst-case reorder buffer size

$$
\begin{aligned}
S_{rb}^{max} &= \{R_m^*(t) - R_1^*(s)\} + \{R_m^*(t) - R_2^*(s)\} \\
&\cdots + \{R_m^*(t) - R_n^*(s)\} \\
&= \alpha_1^*(\Delta t_1) + \alpha_2^*(\Delta t_2) \cdots + \alpha_n^*(\Delta t_n) \\
&= \{r_1(\Delta t_1) + b_1 + r_1 T_1\} \\
&+ \{r_2(\Delta t_2) + b_2 + r_2 T_2\} \\
&\cdots + \{r_n(\Delta t_n) + b_n + r_n T_n\} \\
&= \{r_1\left(h(\alpha_m, \beta_m) - D_1^{low}\right) + b_1 + r_1 T_1\} \\
&+ \{r_2\left(h(\alpha_m, \beta_m) - D_2^{low}\right) + b_2 + r_2 T_2\} \\
&\cdots + \{r_n\left(h(\alpha_m, \beta_m) - D_n^{low}\right) + b_n + r_n T_n\} \\
&= \{r_1(T_m + \frac{b_m}{R_m} - D_1^{low}) + b_1 + r_1 T_1\} \\
&+ \{r_2(T_m + \frac{b_m}{R_m} - D_2^{low}) + b_2 + r_2 T_2\} \\
&\cdots + \{r_n(T_m + \frac{b_m}{R_m} - D_n^{low}) + b_n + r_n T_n\}
\end{aligned}
$$

Worst-case Size Reduction of Reorder Buffer

(1) Adjacency Matrix:

$$A = (a_{ij})_{v \times 6}$$

$$a_{ij} = \begin{cases} p_{ij}, & \text{splitting propotion in } j \text{ direction at node } v_i \\ 0, & \text{else} \end{cases}$$

(2)Contention Matrix:

flow rate and burstiness of sub-flow i

$$\begin{aligned} C_{s_i,d_i} = & [(\varepsilon r_1 + \beta b_1) A_{s_1,d_1} + (\varepsilon r_2 + \beta b_2) A_{s_2,d_2} + \cdots \\ & + (\varepsilon r_j + \beta b_j) A_{s_j,d_j} + \cdots \\ & + (\varepsilon r_k + \beta b_k) A_{s_k,d_k}] \wedge A_{s_i,d_i}, j \neq i \end{aligned}$$

ε and β are the impact factor of ri and b

Worst-case Size Reduction of Reorder Buffer

(3)Congestion Factor:

the total number of nodes in the sub-flow

the direction next to node i

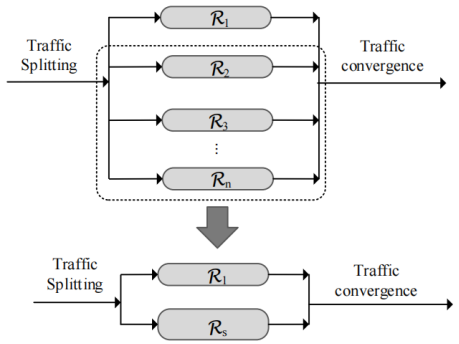
$$\lambda_{f_i} = \sum_{i=1}^{N} C(node\_i, direction\_j)$$

the i-th node along the sub-flow

(4)RB-OLITS:

$$S_{rb}^{max} = \{r_1(T_m + \frac{o_m}{R_m} - D_1^{low}) + b_1 + r_1 T_1\} + \{r_2(T_m + \frac{b_m}{R_m} - D_2^{low}) + b_2 + r_2 T_2\} \cdots + \{r_n(T_m + \frac{b_m}{R_m} - D_n^{low}) + b_n + r_n T_n\}$$

$$\mathcal{R}_s = \sqrt{1/((\frac{1}{\mathcal{R}_2})^2 + (\frac{1}{\mathcal{R}_3})^2 + \cdots (\frac{1}{\mathcal{R}_n})^2)}$$



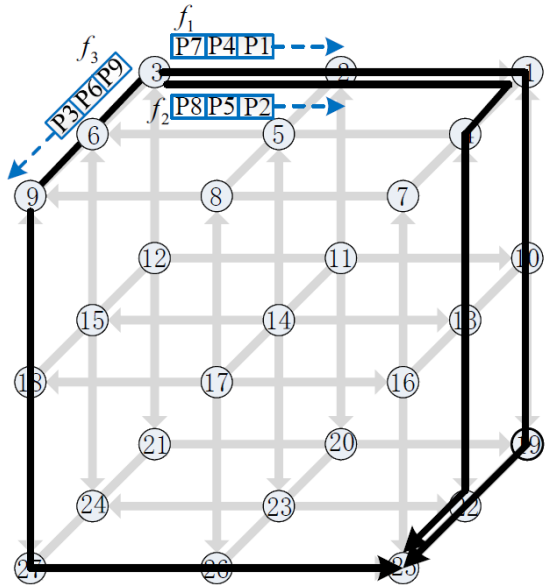
$$\mathcal{P}_{\langle s,d \rangle} = \frac{\mathcal{R}_s^2}{\mathcal{R}_1^2 + \mathcal{R}_s^2}$$

Outline

- Background and Motivation
- Preliminaries
- Worst-case Size of Reorder Buffer
- **Experiments and Results**
- Conclusion

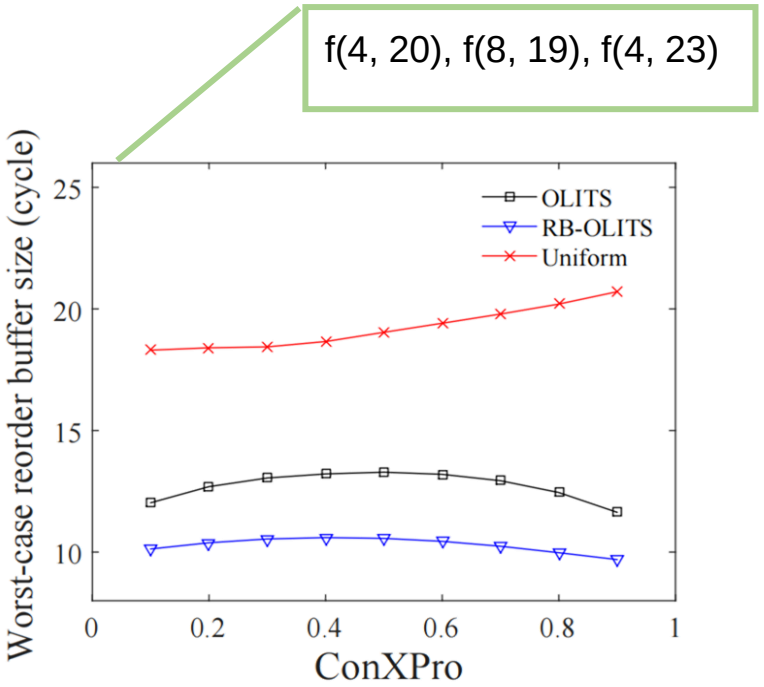
Experiment and Result

Synthetic Pattern



tag flow

f(3, 25)

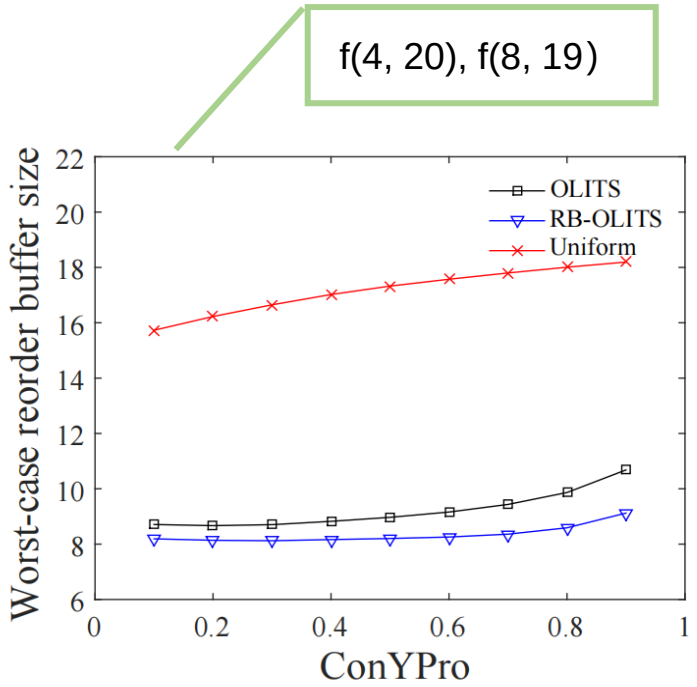


vs. Uniform

max: 53.20% average: 46.35%

vs. OLITS

max: 20.88% average: 19.09%

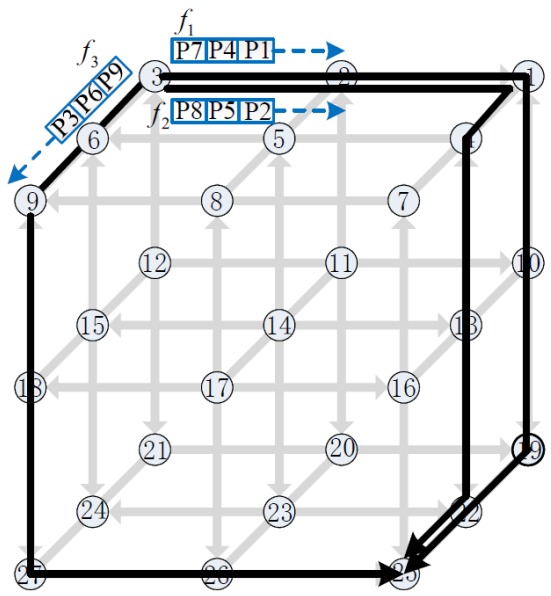


vs. OLITS

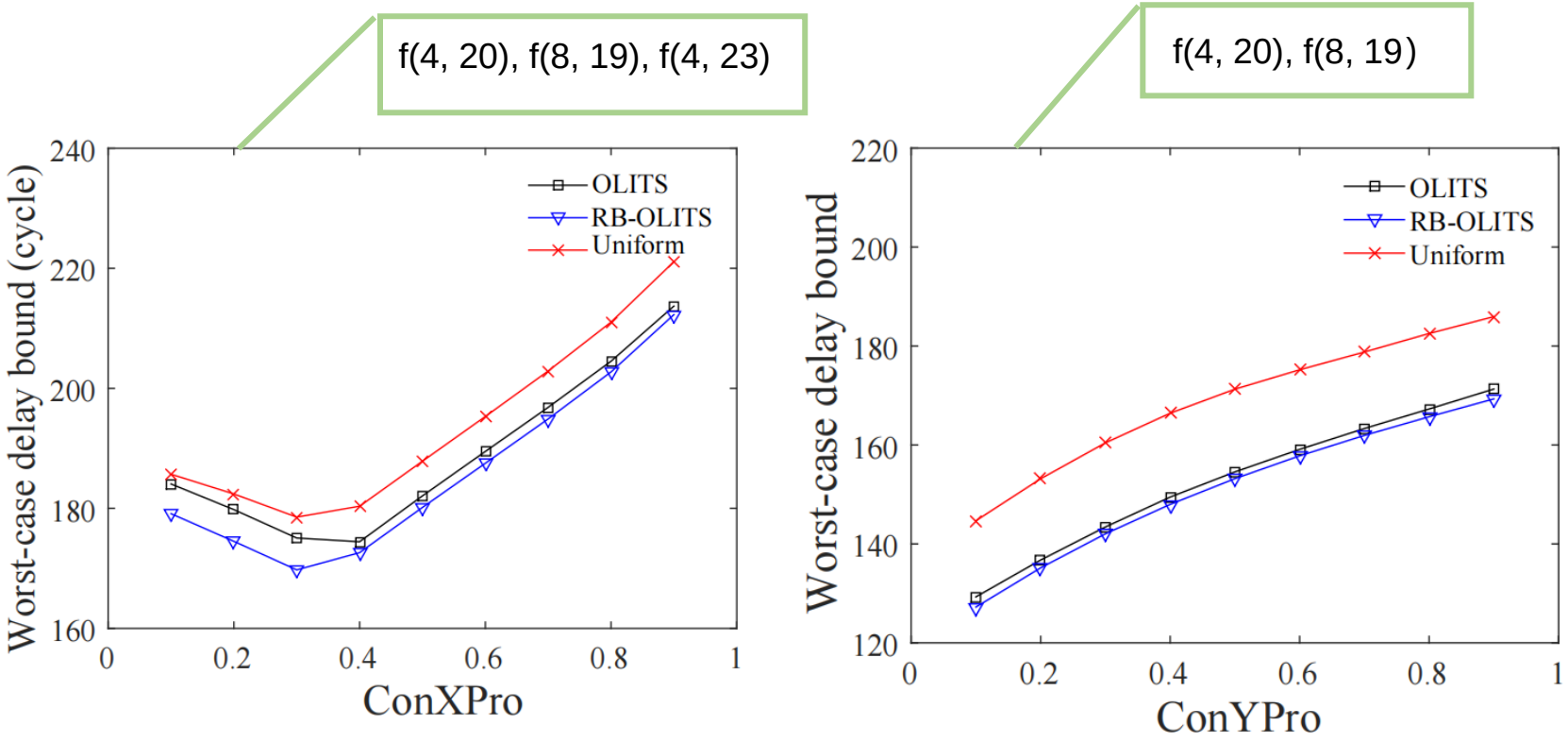
max: 14.60% average: 9.28%

Experiment and Result

Synthetic Pattern



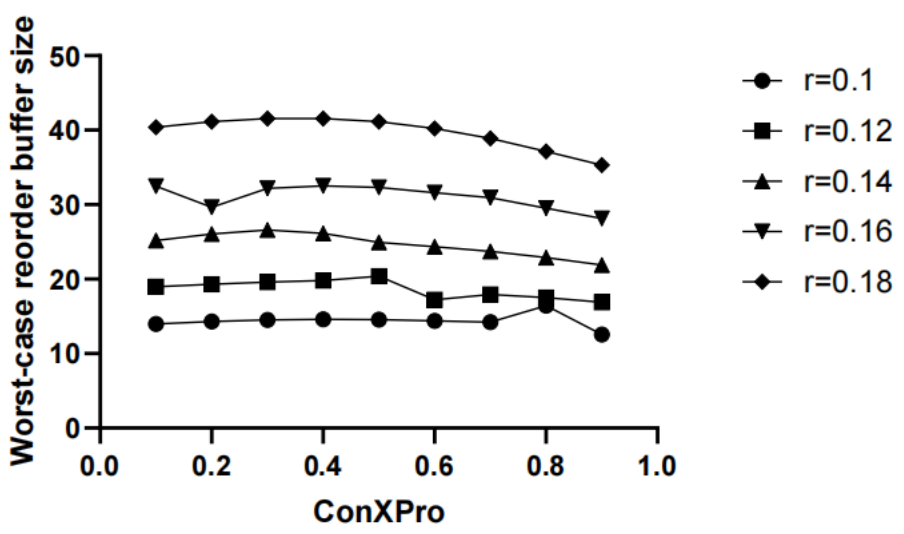
tag flow
f(3, 25)

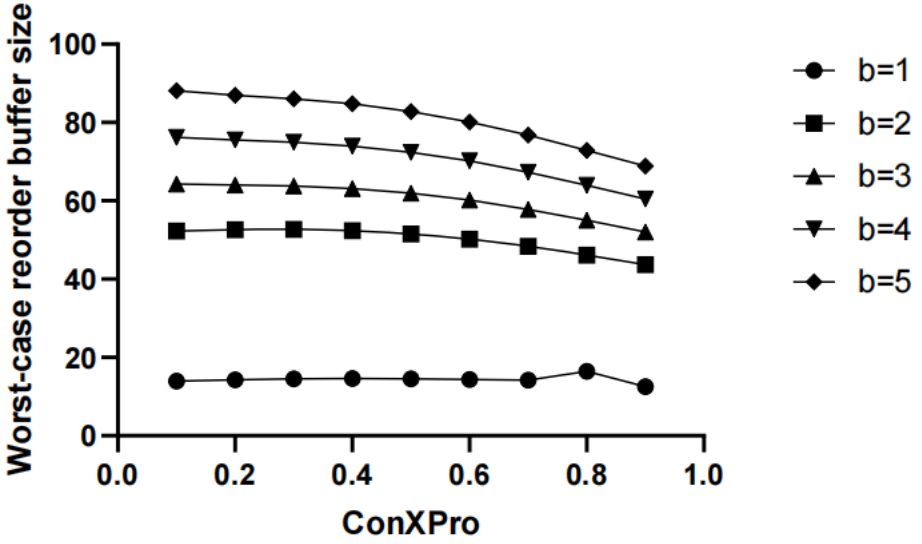


Experiment and Result

Synthetic Pattern:

Impact of injection rate with varying bustiness

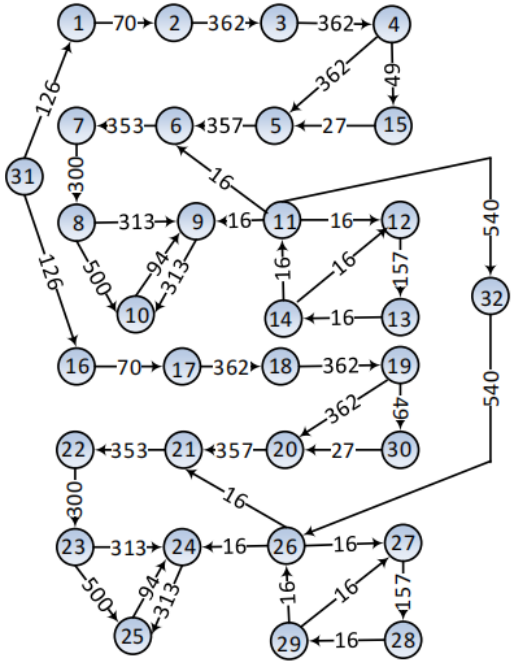




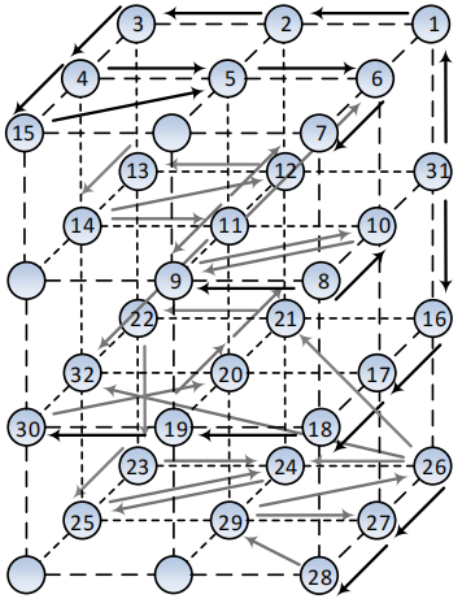
With the increasing injection rate or burstiness, the worst-case reorder buffer size increases accordingly

Experiment and Result

Industrial Pattern



(a) Task Graph

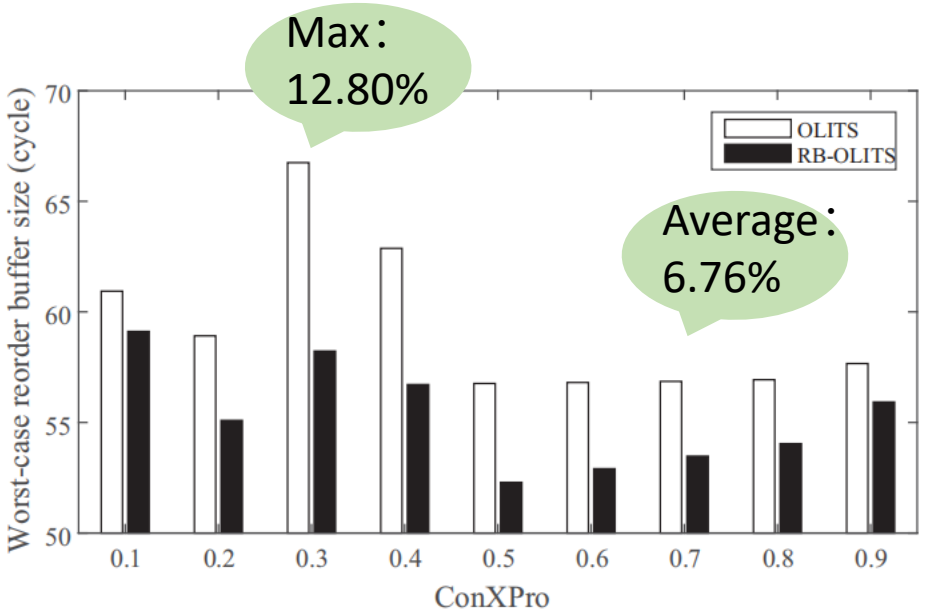


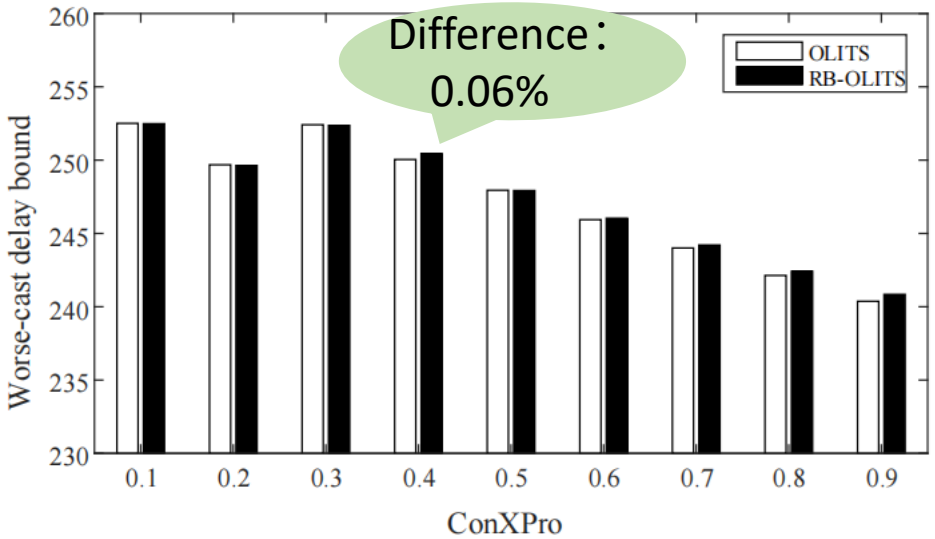
(b) Task Mapping

43 contention flows

$\beta(t) = 0.33 \cdot (t - 3)^{+}$

12 sub flows





Outline

- Background
- Preliminaries
- Worst-case Size Reduction of Reorder Buffer
- Experiments and Results
- **Conclusion**

Conlusion

■ **An analytical model for calculating the worst-case reorder buffer size for multi-path minimal routing 3D-NoCs using network calculus**

- support arbitrary number of sub-flows

■ **A traffic splitting method, named RB-OLITS, for reducing worst case reorder buffer size.**

- an improvement of 5.9%-20.9% in most cases

Thank you!