A^{2}OLPC

Performance Impact of SMP-Cluster on the On-chip Large-scale Parallel Computing Architecture

Shenggang Chen, Shuming Chen, and Yaming Yin

National University of Defense Technology
Changsha, China

{shgchen,smchen,ymyin}@nudt.edu.cn

PMEO-2010

Table of Contents

- Introduction
- Performance Models
- Simulation Results and Discussion
- Experiments
- Conclusion and Future Work

Introduction

- the computing paradigm of the massive parallel computing systems is being applied to the on-chip systems
    - one of the examples is the manycore design methodology, also called On-chip Large-scale Parallel Computing architectures(OLPCs)
- Enhanced core is necessary to the high system performance; however, the parallelism suffers [based on Hill’s research]

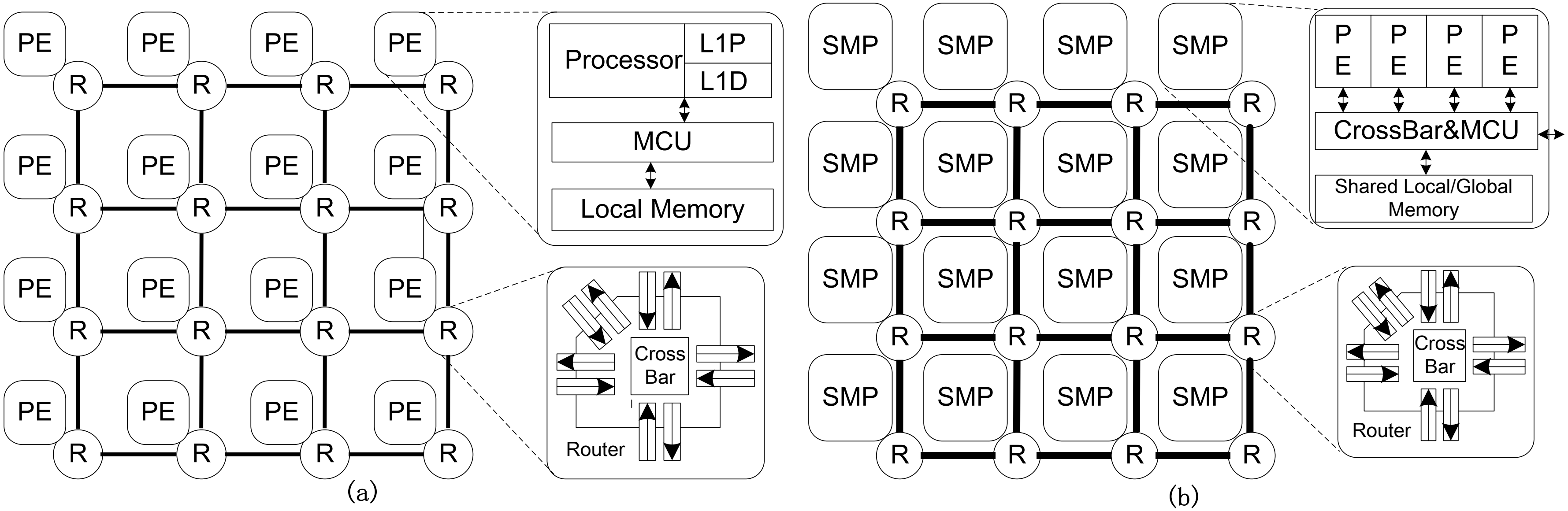
Characteristics of Applications

- Application is the original driven to the architecture improvement
- the parallelisms characteristics in these application are of diversity
  - some parts in the application can be parallelized to much more threads than the other parts, or
    - Processors capable of this are called dynamic multicore and it is still hard for software to control the dynamic resources
  - hierarchy of parallelism can be found in the applications
    - Hierarchical OLPCs are capable of this

Motivation

- Hierarchy of parallelism in application requires the support of hierarchy architectures
- Clustered architectures in massive parallel computing networks are popular for decades.
  - Can it be beneficial when applied to the OLPCs?
- Zooming communication delay will significantly constrain the performance improvement when the PE counts in a OLPC becomes larger
  - How to make the delay of the on-chip interconnection goes up slower?

- To answer these questions, we should make a comparison of different types of OLPCs
- **Tiled-flat architecture**
  - using tiled-flat interconnection for communication, in which each node is a basic Processing Element
- **SMP-clustered architecture**
  - groups multiple PEs to form a shared-memory cluster and connects the clusters with size-reduced on-chip interconnection

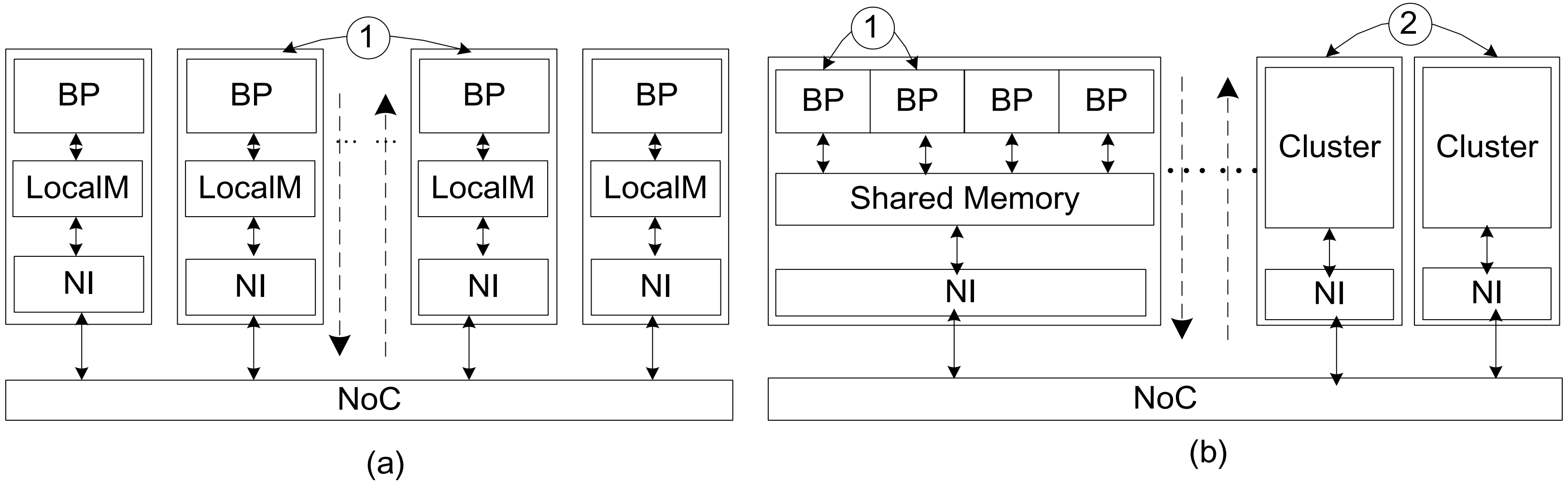


(a)

(b)

- (a) Tiled-flat Arch.
  - **Tiled-flat** OLPCs scale well, but communication delay significantly constrains performance improvement when PE counts becomes large
- (b) SMP-Clustered Arch.
  - **SMP-Clustered** OLPCs will benefit from
    - the same way of clustered architecture in massive parallel computing systems
    - the locality of the communication and the reduction of the intercnnection size with no parallelism degree cost

Communication Patterns



- (a) single-level communication:
  - LongnP/Log3P
- (b) tow-level communications: vertical and horizontal
  - mLognP

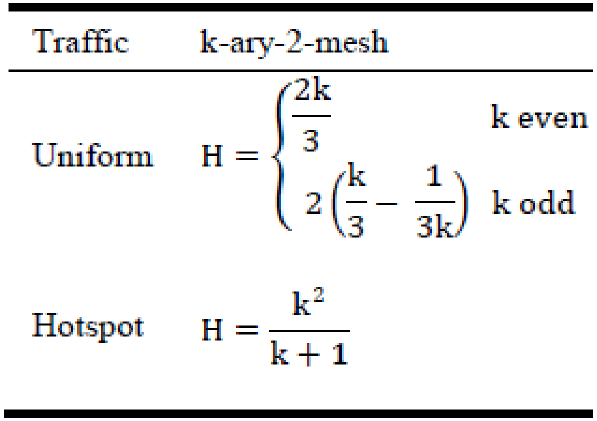
Performance Models

- Basic Assumptions:
  - Serial tasks in an application do not introduce data communication, only parallel tasks do
  - Considering the limitation of the technology, the number of the PEs in a SMP node should not increase randomly. So we believe that the capacity increment of the on-chip memory in a SMP node does not induce the increment of the access latency to more than one cycle. This assumption holds true for the small-scale SMP node. In the future work this assumption will be eliminated.
  - Only k-ary-2-mesh NoC architecture is considered

Networks-on-Chip Delay Model

$$\Gamma_{1P}(N) = H \cdot t_c + \left\lceil \frac{L}{b} \right\rceil$$

$$\Gamma_{1C}(N) = M \cdot \Gamma_{1P} = M \cdot H \cdot t_c$$



Traffic	k-ary-2-mesh
Uniform	$H = \begin{cases} \frac{2k}{3} & \text{k even} \\ 2\left(\frac{k}{3} - \frac{1}{3k}\right) & \text{k odd} \end{cases}$
Hotspot	$H = \frac{k^2}{k+1}$

Communication/Computation Ratio

- Communication frequency/computation

$$f(N) \sim \gamma \cdot \sqrt{N}$$

  - γ is a constant that determines the degree of relationship

- Tiled-flat: $\Gamma_p \cdot f(N)$
- SMP-Clustered:
  - Inter-node: $\Gamma_p \cdot f\left(\frac{N}{C}\right)$
  - Intra-node: $\Gamma_p \cdot \left(f(N) - f\left(\frac{N}{C}\right)\right)$

Speedup Models

$$S = \frac{(\Gamma_s + \Gamma_p)}{\left(\Gamma_s + \frac{\Gamma_p}{N}\right) + \Gamma_o}$$

$$S_{\text{flat}} = \frac{1}{\left(1 - \beta + \frac{\beta}{N}\right) + \beta \cdot \gamma \cdot \sqrt{N} \cdot \Gamma_{1C}(N)} \qquad \beta = \frac{\Gamma_p}{\Gamma_s + \Gamma_p}$$

$$S_{\text{hier}} = \frac{1}{\left(\frac{1-\beta}{\sigma} + \frac{\beta}{N}\right) + \beta \cdot \gamma \cdot \sqrt{N} \cdot \left(\frac{1}{\sqrt{C}} \cdot \Gamma_{1C}\left(\frac{N}{C}\right) + \left(1 - \frac{1}{\sqrt{C}}\right) \cdot g\right)}$$

$$\Gamma_{1C}\left(\frac{N}{C}\right) = M \cdot H\left(\frac{N}{C}\right) \cdot t_c \qquad H\left(\frac{N}{C}\right) \approx \frac{H(N)}{\sqrt{C}}$$

Winning Condition

$$(\sigma - 1)\frac{(1-\beta)}{\sigma} + \beta\gamma\sqrt{N}\left(\Gamma_{1C}(N) - \left(\Gamma_{1C}\left(\frac{N}{C}\right)\frac{1}{\sqrt{C}} + g\left(1 - \frac{1}{\sqrt{C}}\right)\right)\right) > 0$$

consider $(\sigma - 1)\frac{(1-\beta)}{\sigma} > 0$, and $(\sigma - 1)\frac{(1-\beta)}{\sigma} \to 0$, $H\left(\frac{N}{C}\right) \approx \frac{H(N)}{\sqrt{C}}$

$$\mathbf{g \leq H(N) \cdot M \cdot t_c \cdot \left(1 + \frac{1}{\sqrt{C}}\right)}$$

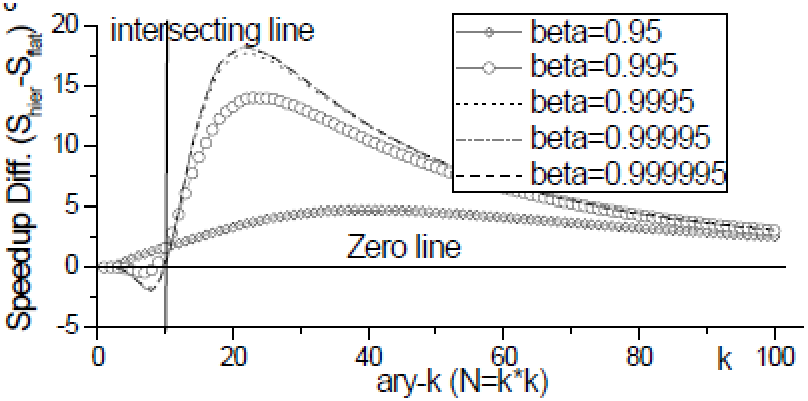
- If the inequality is satisfied, the SMP-clustered OLPCs will outperform the Tiled-flat OLPCs.
  - H(N) is a function determined by the basic PE count and the traffic model. tc is architecture dependent and g will slowly increase with the increment of C, i.e., the capacity of the shared memory, which is omitted in our analysis.
  - For a certain N, increment of C will induce the increment of g, meanwhile the right side of the inequality will get a reduction. Therefore, there should be a balanced point for the value of C to achieve an optimal speedup

Simulation and Discussion

- The experiential values of the parameters are employed in the performance models
- Matlab
- Contents:
  - Performance comparison
  - Impact of PE Count: N
  - Impact of Collective Degree: C
  - Impact of the Acceleration of SMP to the Serialization Program: σ

Performance Comparison

- **Phenomenon 1.** When the size of the OLPC system becomes a certain extent large, SMP-clustered architectures will certainly outperform the Tiled-flat architecture in terms of speedup.
- **Phenomenon 2.** Some negative values appear at the low extreme of x-axis. The Tiled-flat architecture outperforms the SMP-clustered architecture. This is because for small-sized network, the hop count is small and the SMP shared memory access latency is relatively too long to keep the winning condition satisfied.



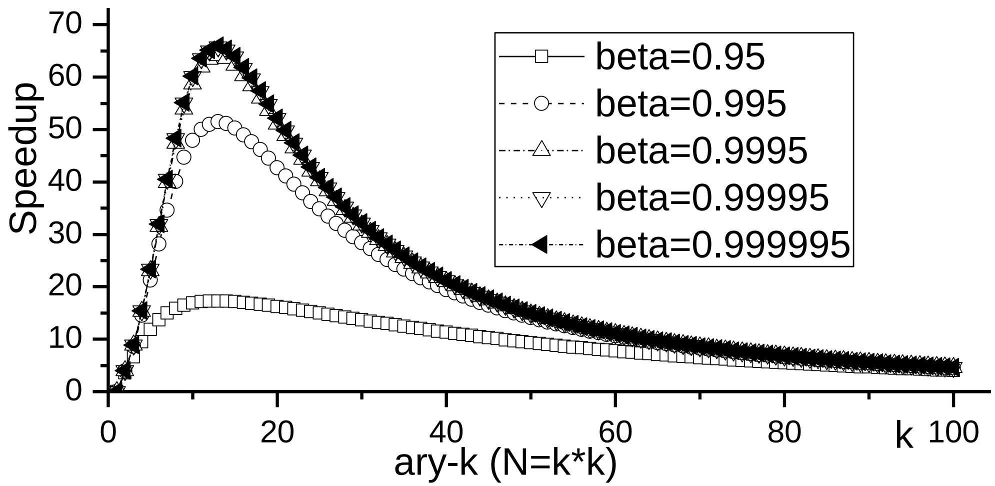
Phenomenon 3. The bigger the ratio of the parallel part to the whole application is, the more speedup advantage the SMP-clustered OLPC can get.

Implication 1. The parallelism of application is at the top of importance. Programmer should never stop to exploit more parallelism and parallel part of the application. This is not new and has been published in Hill’s work[3].

•**Phenomenon 4.** As the size of OLPC becomes large enough, the advantage of the SMP-clustered architecture fades out. This is because the value of C is too small to reduce the hop count of the large-scale network.

•**Implication 2.** As the number of the PEs in a OLPC increases, one should gradually increase the value of the collective degree parameter C to keep the advantage of the SMP-clustered architecture.

Impact of PE Count: N

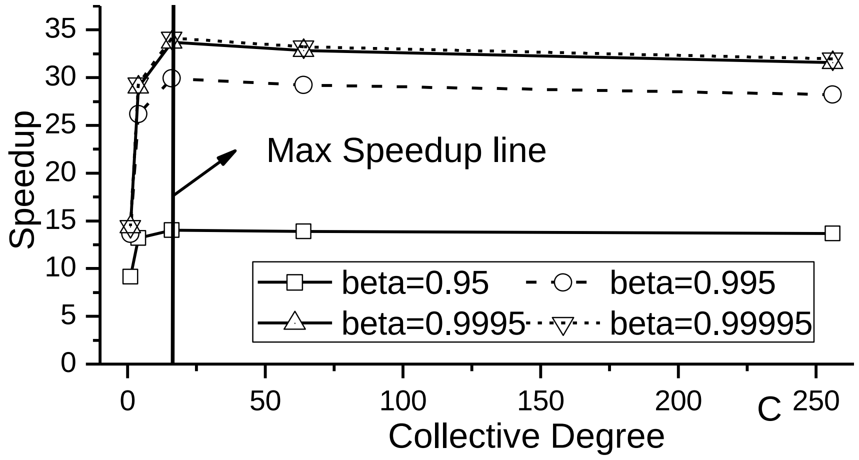


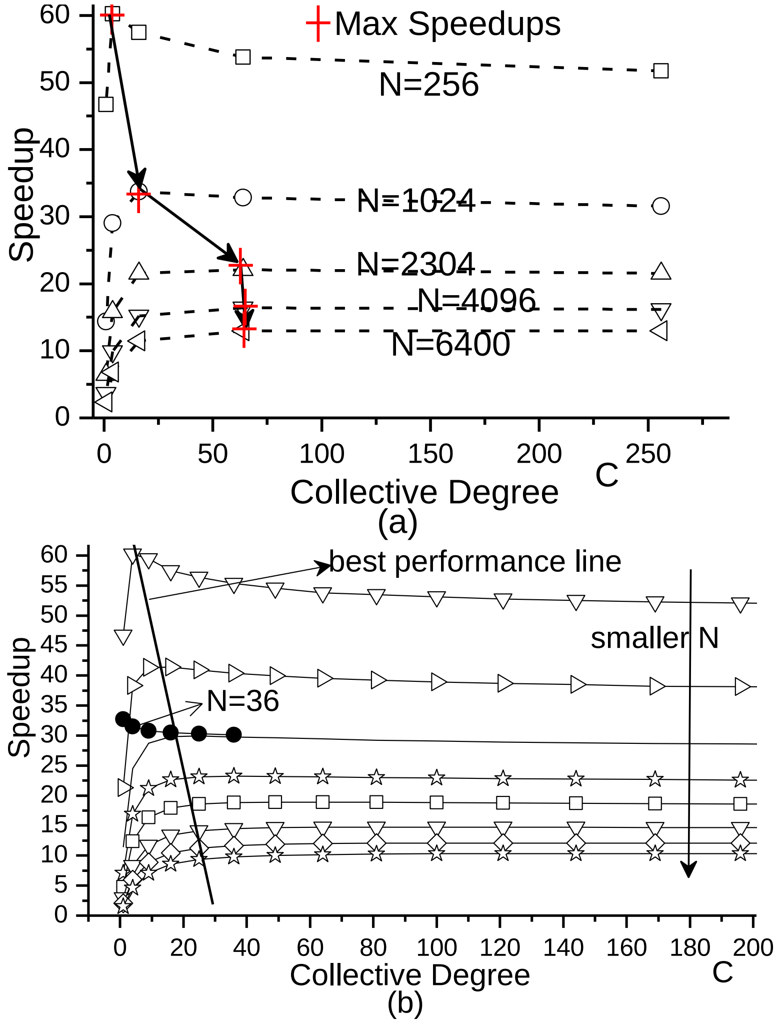
Here we can still observe the impact of the application character, i.e., the parallel ratio. The larger is, the cliffier the upside of the hump is and the bigger the magnitude of the hump is. As the ratio increases, the peak of the hump shifts right.

- **Phenomenon 5.** As the PE count N increases, all curves have a hump with a cliffy upside and a slow downside. When the PE count N is large enough, the speedup of the system becomes very small because the benefit of the large parallelism is amortized by the increased network latency.
  - **Implication 3.** More efficient interconnection networks should be invented. When the PE count becomes large enough, the network latency becomes one of the shackles for performance speedup. Luckily, nowadays OLPCs still keep in the K-level (thousand) and the problems are not so urgent.
  - **Implication 4.** The increment of the parallel hardware resources should be with the increment with the problem size. For fixed problem size, more available parallel resources may cause the reduction of the performance speedup or a waste of the resources.

Impact of Collective Degree:C

- let $\frac{\partial S_{hier}}{\partial C} = 0$, we get $C_{opt} = \left[\left(\frac{4M\sqrt{N}t_c}{3g}\right)^2\right]$
    - increases with the increment of N, decreases with increment of memory access latency g
- At the beginning, the latency of the shared memory access is smaller than that of NoCs, so that the performance speedup increases with the increment of C. After crossing the optimal value, the latency of NoC is getting smaller and collecting more PEs will gain no more profit.



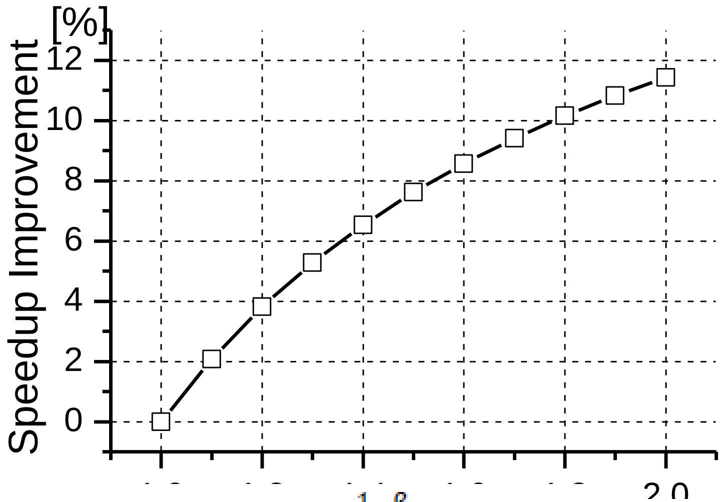


(a)

(b)

- **Implication 5.** Taking the implementation into consideration, there is no significant benefit to continue grouping more PEs even when the ideal optimal value is not exceeded. C=5x5 will be good for most of the OLPCs with thousands of PEs

Impact of the Acceleration of SMP to the Serialization Program: σ



At present, there are some directions targeting at this, such as data speculation[14], thread-level speculation[13], hardware dynamic recomposing[5-6] and so on. However, it is still a tough work both of the architecture designers and the compiler researchers. Chips with this ability belong to the dynamic multicore chips as stated by Hill[3]. Considering these difficulties, we believe that the SMP-clustered architecture is a promising tradeoff between performance and the implementation cost

In model S_{hier} , item $\frac{1-\beta}{\sigma}$ reveals the acceleration of the SMP cluster to the serialization program.

- **Implication:** Researchers should continue to study the acceleration techniques of the single-thread program. This is the same as published by Hill

Conclusion and Future Work

- Key contributions:
  - The model for the SMP-clustered architecture is obtained by extending of Amdahl's Law, and the comparison between the SMP-clustered architecture and the Tiled-flat architecture is conducted through model analysis and experiential values of the parameters used in the speedup model. Analytical condition for SMP-Clustered architecture's advantage is also drawn. The conclusions of this paper support the communication paradigm using locally symmetrically-shared memory, and it is efficient to slow down the performance reduction brought in by on-chip interconnection.
  - The design space of the SMP-clustered architecture is exploited, especially for the collective degree. Some instructive conclusions for architecture design and application development are obtained through analysis of the model influence factors.

- the models do not consider the contention for communication and the SMP shared memory. The contention of the communication can significantly increase the latency and thus reduce the speedups. The data sharing among multiple threads through shared memory will induce application-dependent performance reduction. For more realistic consideration, the contention should be added in.
- The further proving of the model should be conducted

- Thanks!