Milestone2

Rachata Ausavarungnirun, Kevin Chang
Overall problem and idea

- Applications are slowed down differently based on each applications’ characteristics
- Throttling back some applications can improve system throughput
- Can we make a system that can deliver high throughput and fairness at the same time?
Algorithm 1

Global Round-Robin Scheduling

- How to determine congestion?
  - Use average network MPKI as a metric

- When to throttle?
  - Sample the network state every 20k cycles.

- Whom to throttle?
  - Applications with high MPKI.

- How to throttle?
  - Randomly distribute applications with high MPKI to three clusters
  - Each cluster permitted to inject freely for a certain interval
  - While one cluster is injecting freely, other clusters are being throttled
Algorithm 2

Global RR + Batch Scheduling

- How to determine congestion?
  - Use network utilization as a metric

- When to throttle?
  - Sample the network state every 20k cycles.

- Whom to throttle?
  - Applications with high MPKI.

- How to throttle?
  - Distribute high MPKI applications to form clusters with equivalent total MPKI
  - Each cluster permitted to inject freely for a certain interval
  - While one cluster is injecting freely, other clusters are being throttled
Algorithm 3

**Algorithm 2+ network-aware application selection**

- **How to determine congestion?**
  - Use network utilization as a metric

- **When to throttle?**
  - Sample the network state every 20k cycles.

- **Whom to throttle?**
  - A fraction of applications with high MPKI.

- **How to throttle?**
  - Distribute high MPKI applications to form clusters with equivalent total MPKI
  - Each cluster permitted to inject freely for a certain interval
  - While one cluster is injecting freely, other clusters are being throttled
Results

Minimal gain

Benefit the most

Not congested
Results

Minimal gain

Not congested

Benefit the most

Workload IPC by Algo

IPC

Workloads
Current status

- Collecting results on fairness
- Implementing the mechanism proposed in Nychis et al, “Next Generation On-Chip Networks: What Kind of Congestion Control Do We Need?”
- Mechanism to identify applications to be throttled.
- Mechanism to perform more aggressive throttling.
  - Dynamically adjust the throttle rate.
  - Dynamically adjust the MPKI batch limit for each cluster.