Coordinated Scheduling: A Mechanism for Efficient Multi-Node Communication

Edward W. Knightly and Chengzhi Li
Rice Networks Group

http://www.ece.rice.edu/networks
Background: Priority Scheduling

- Each packet has a priority index

- Scheduler selects smallest priority index pkt first

- Index assignment scheme ⇒ Service Discipline
  - **FIFO**: index = arrival_time
  - **Virtual Clock**: index = max(arrival_time, prev_index + L/r)
Earliest Deadline First

- Scheduler services packet with smallest deadline = arrival_time + delay_bound

- EDF is optimal for a single server
Multiple Nodes: Issue 1, Sub-Optimality

- Over multiple nodes, EDF is **not** optimal
  - Locally optimal rules do not achieve global optimum (best end-to-end performance)

⇒ ... Can do better
Multiple Nodes: Issue 2, Traffic Distortion

- Traffic can become more bursty downstream
  - Arrivals previously in $[t-d, t+I]$ now in $[t, t+I]$

- Consequence: difficult to analyze and efficiently support multi-node QoS
Existing Solutions to Distortion Problem

1. **Reshape traffic**
   Hold packets until conform to original pattern

2. **Isolate flows**
   Limit distortion by limiting sharing (e.g., guaranteed rate)

- **Problems**
  - **Utilization** impact of isolation/non-work-conserving
  - **Scalability** issues with per-flow operations

Edward W. Knightly
Grand Challenge

Design a scheduler with the following properties

- **Efficient**
  - achieves high utilization and is work-conserving

- **Scalable**
  - without per-flow mechanisms

- **Quality of Service**
  - Provides mechanisms for end-to-end services
Our Approach: Coordination

- Virtual coordination among servers
  - Router computes priority index as a function of upstream index

- Implications
  - Late packets upstream have increased priority downstream
  - Early packets have priorities reduced downstream
Remaining Outline

- Devise a general **framework & definition** for coordination
- Show that **CEDF, FIFO+, CJVC, ...** belong to the CNS class
- Derive **end-to-end** schedulability conditions of CNS networks
  - results apply to all schedulers
- Illustrate **performance** implications of coordination
Coordinated Network Scheduling Definition

- CNS is a work conserving scheduler that selects the packet with the smallest priority index first.

- Indexes are given by:

\[ d_{i,j}^k = \begin{cases} 
    t_i^k + d_{i,1}^k & \text{at the first hop} \\
    d_{i,j-1}^k + d_{i,j}^k & \text{at the } j^{th} \text{ hop} 
\end{cases} \]

\( d_{i,j}^k \) = priority index of the \( k^{th} \) packet of flow \( i \) at its \( j^{th} \) hop

\( t_i^k \) = (virtual) arrival time of the \( k^{th} \) packet of flow \( i \) at the first hop

\( d_{i,j}^k \) = the increment of priority index of the \( k^{th} \) packet at the \( j^{th} \) hop

- Observe the recursive relationship of priorities, i.e., coordination
Coordinated Network Scheduling

- **Observation:** A number of (old and new) schedulers employ coordination
  - Recursive priority index

- **Goal:** Identify their common elements and study the class under a single framework
FIFO+ [CSZ92]

- Servers measure $d$, the average local queueing delay, and actual packet delay $\hat{d}$.
- First node is FIFO.
- Downstream priority index is accumulated $d - \hat{d}$ terms from upstream nodes.
- Multi-node performance gains over WFQ [CSZ92].

Edward W. Knightly
FIFO+ is a Coordinated Scheduler

- Specifying scheduler is CNS index assignment

\[ d_{i,1}^k = 0 \rightarrow \text{FIFO at first hop} \]
\[ d_{i,j}^k = \bar{d}_{i,j-1}^k - \hat{d}_{i,j-1}^k \rightarrow \text{Downstream, relative delay is accumulated, and adjusts priority} \]
Coordinated Earliest Deadline First (Similar to [And99,CWM89])

- CEDF uses virtual coordination among servers
  - Downstream priority index is a function of upstream index \((t+5+5\text{ vs. } u+5)\)

- Late packets upstream have increased priority downstream
  - Ex. Pkt delayed by 9 has 2\(^{nd}\) node index 1 (vs. 5)

- Early packets have priorities reduced downstream
  - Ex. Pkt delayed by 1 has 2\(^{nd}\) node index 9 (vs. 5)
Core-stateless Jitter-controlled Virtual Clock (CJVC) [SZ99]

- CJVC’s goal: per-flow QoS guarantees without per-flow state in the core
  - Mechanism: Dynamic Packet State (DPS)

- Observe: CJVC has recursive priority among nodes
  - CJVC $\in$ CNS

Edward W. Knightly
CNS Properties

- All CNS schedulers are core-stateless and scalable

- CJVC, FIFO+, ... can be viewed as CNS index assignment schemes
  - Rate-CNS
    - priority index depends on reserved bandwidth (ex. CJVC)
  - Delay-CNS
    - index depends on delay parameter (CEDF, FIFO+, OCF)
Advantage of CNS Framework

- Improved understanding of multi-node mechanisms

- Scheduler design
  - CEDF: end-to-end delay bounds
  - CJVC refinement: work-conserving and without “slack variable”

- Performance analysis and QoS
  - Solve CNS, solve all...
Theoretical Results

- **Essential Traffic Envelope (ETE)**
  - Traffic interfering with ability to meet QoS target

- **Bound ETE downstream**
  - Exploit coordination property
  - Prove distortion limited, much as with reshapers

- **Bound end-to-end delay**
  - Local (per-node) violations permissible

- **Index assignment schemes**
  - CNS can achieve delay bounds of WFQ
Traffic Envelopes

- Envelopes characterize arrivals as a function of interval length
  - Max and deterministic [Cr95, KWLZ95]
  - Statistical [QK99]

- Recall: traffic distortion problem
  $\Rightarrow$ envelopes distorted

$E^*(I) = 3$
New Concept: Essential Traffic Envelope

- Essential traffic impedes a packet’s ability to meet a deadline
  - Ex. with FIFO, it’s pkts arriving earlier
- Approach: bound traffic with a deadline range vs. an arrival time range (ETE vs. TE)
Illustration: First Hop (EDF and CNS)

- 1st hop: priority indexes are the same in CNS and EDF
- Suppose that the third packet is seriously delayed due to cross traffic
At the second hop, the priority indexes depend on the (local/late) arrival times in EDF.

Traffic distortion is large and propagates downstream.
2\textsuperscript{nd} hop: the priority indexes are independent of the (local/late) arrival times in CNS

- Departures are narrowly distorted (\textit{without} reshaping)
- Theory tightly bounds distortion of essential traffic
End-to-End Schedulability Condition

- Allow local violations (ex. missed per-node deadlines)
  - ...contrast to all previous work

- Bound Essential Traffic Envelope downstream

- Derive an end-to-end delay bound

Schedulability Condition for all coordinated schedulers (CEDF, CJVC, GEDF, FIFO+, ...)
  - CEDF, GEDF, ... not previously derived
  - CJVC bound tighter than [ZDH01]
Index Assignment

- Recall: indexes can be delay targets or L/r rate assignments
- Result: under CJVC-like rate assignment and leaky bucket constrained flows

Coordinated scheduling achieves the same end-to-end delay bound as WFQ

⇒ Same WFQ bounds, yet scalable, work conserving, ...

⇒ CNS is no worse than WFQ. But can be much better!
Performance Analysis: CNS vs. GPS

- Two CNS weight assignment schemes:
  - S-CNS (Simplified CNS)
    - Constant local delay assignment scheme (2 and 6 msec respectively)
  - G-EDF (Global EDF) [CWM89]
    - Uniform allocation with larger weight at first node
**Voice Flows 64/32 kb/sec**

- **Advantages of coordination**
  - lower end-to-end delay bounds and larger admissible regions

Edward W. Knightly
CNS vs. EDF (Pareto on-off)

- With 300 flows, reduction in delay from 120 msec to 50 msec
CNS vs. WFQ

- With 300 flows, delay reduced from 170 to 50 msec
Conclusions

- CNS provides a framework for coordinated and scalable schedulers
  - FIFO+, CJVC, GEDF, CEDF, ...
- General end-to-end results for CNS class
  - Bound downstream envelopes exploiting recursive priority index
- CNS performance advantages
  - Can outperform WFQ, EDF, and re-shaping EDF
RNG Projects

- **Coordinated Scheduling** [LK00,LK01,...]
  - Robustness to parameter allocation
  - Multi-hop wireless networks

- **Web Server and End System QoS** [KK00]

- **Scalable QoS**
  - Edge [CK00, SSYK01] and Host [BKSSZ00] controlled services

- **Multi-class services**
  - Theory [QK99] and measurement [KK01]