LOW POWER NANOSCALE BUFFER MANAGEMENT FOR NETWORK ON CHIP ROUTERS

- Suman K. Mandal, Texas A&M University
- Ron Denton, Texas A&M University
- Saraju P. Mohanty, University of North Texas
- Rabi N. Mahapatra, Texas A&M University

Contact: skmandal@cse.tamu.edu

Acknowledgement: This research is supported in part by NSF award number 0509483 and 0854182.

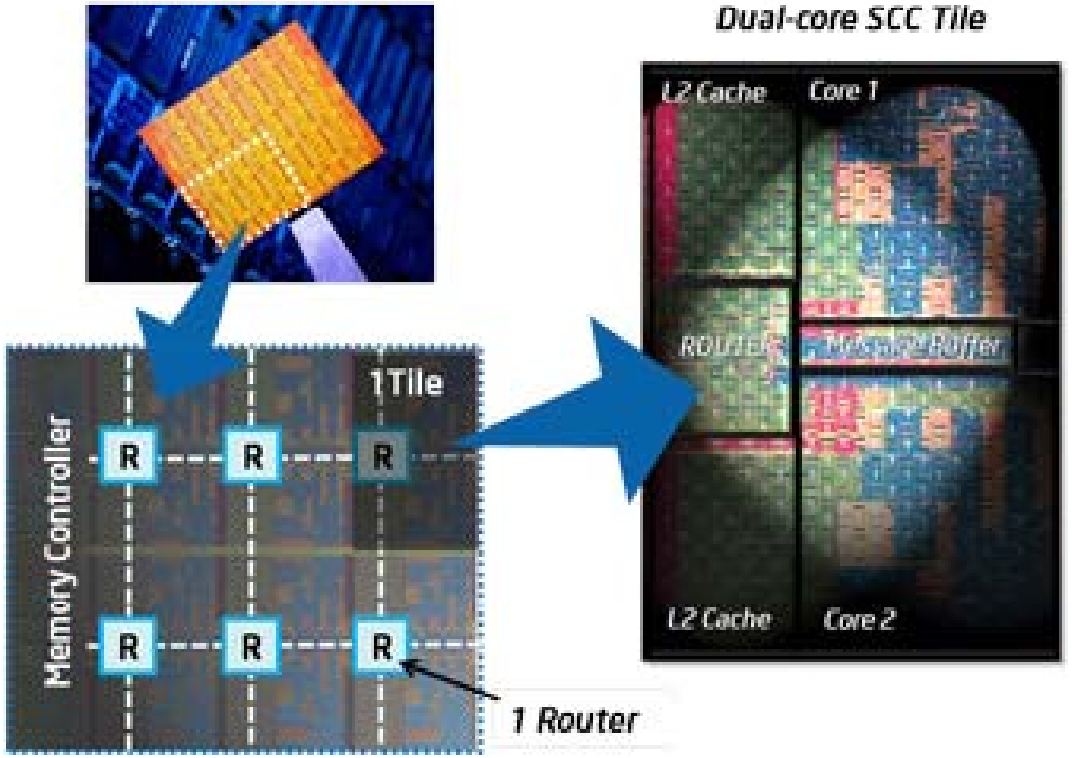
Talk Outline

- Introduction and Motivation
- Contributions
- Related Prior Research
- Router Architecture
- Block-Level Power Management
- Low Power SRAM Buffer
- Flit-Level Power Management
- Conclusions

Introduction and Motivation

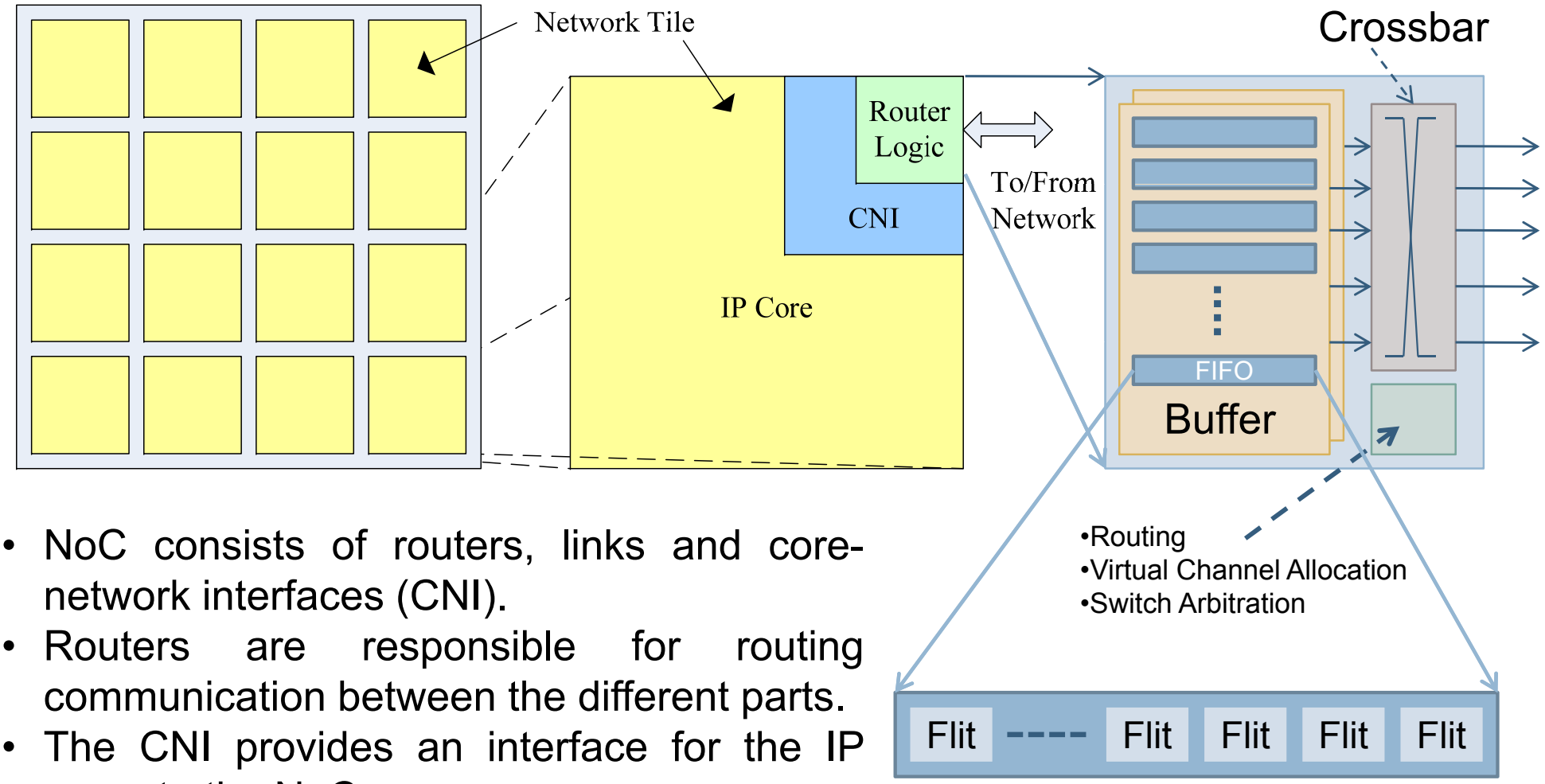
Introduction: Why NoC?

- Network on Chip
  - Next Gen Interconnect
  - GALS Approach
- Advantages
  - High Bandwidth
  - Scalable
  - Extensible
- Disadvantages
  - Power Hungry



48 Core Chip from Intel Uses 24 Node Mesh Network on Chip

NoC: Structure

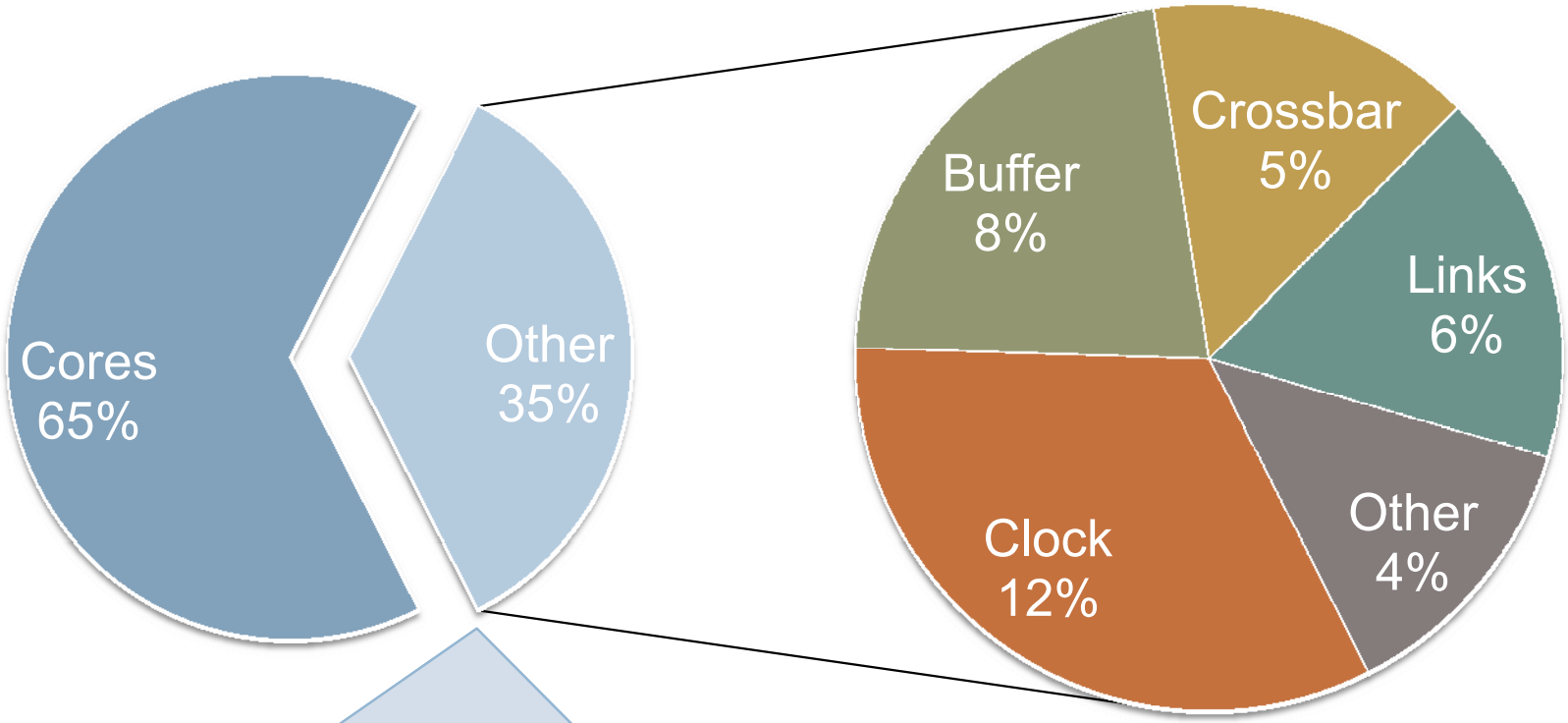


- NoC consists of routers, links and core-network interfaces (CNI).
- Routers are responsible for routing communication between the different parts.
- The CNI provides an interface for the IP cores to the NoC.

NoC: Salient Features

- Buffer size, type and allocation policy play an important role in the performance and efficiency of a NoC router.
- Buffers can consume as much as up to 79% of NoC router power.
- Buffer utilization in NoC router is dependent on network congestion.
- Depending on communication pattern of an application a buffer utilization of a router varies over time.

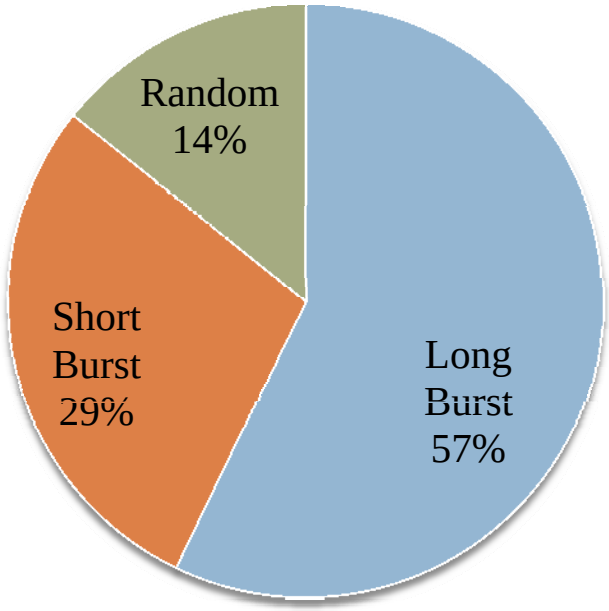
Chip Power Breakdown



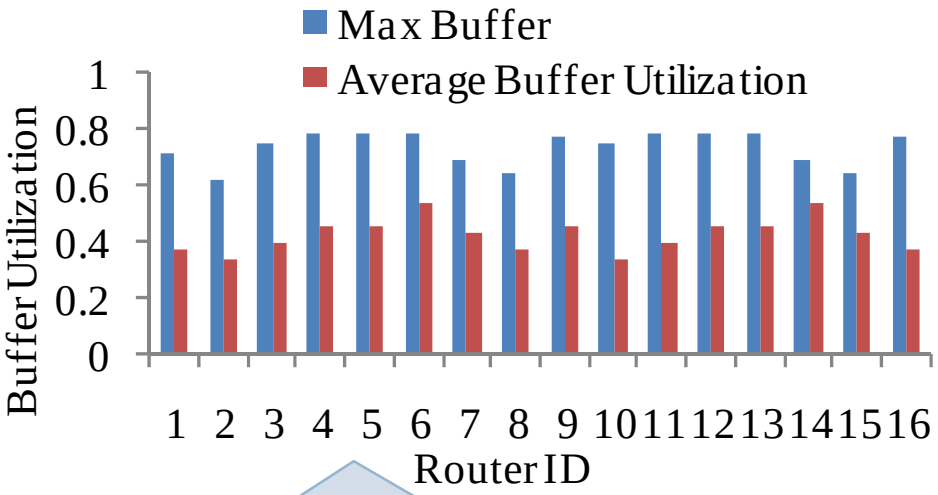
NoC Consumes about 35%
Out of which 22% is buffer

Traffic and Buffer Utilization

Traffic Types By Volume



Buffer Utilization by Position



Although Peak utilization is high, Average utilization is much lower.

Types of NoC Buffers

- First-In-First-Out (FIFO) registers.
- Static random access memory (SRAM) based buffers.

NoC Buffers: SRAM Advantages

- Nanoscale SRAM buffers are suitable for NoC router design because of their speed, density and reliability.

Motivation for this Research

- Efficient buffer management is necessary to ensure high performance and low power.
- Power dissipation characteristics of nanoscale SRAMs are unique and hence traditional low power design techniques are not sufficient to ensure minimum power operation.

Contributions of this paper

Idea!

- Can the knowledge of traffic be utilized to minimize the buffer requirement?
  - Yes, because burst modes can be detected easily.

- How to minimize the buffer requirement?
  - By dynamically resizing the buffers to required size.

The Contributions

- A feedback controlled block-level buffer management is proposed for power management.
- An adaptive controller for efficient flit-level power management is proposed.
- Both power management techniques are thoroughly evaluated for performance.
- Results outperform static allocation by 21% increase in throughput and 20% reduction in energy consumption.

Related Prior Research

Prior Research

- There have been significant research on router buffer power management for low power.
- Both circuit level and system level techniques have been proposed for NoC power management.
- Detail discussion on existing research is available in Simunic-DATE2002, Banerjee-NoCSymposium2007, Ogras-CODES2005.

Prior Research …

- Zhang et al. GLSVLSI 2009:
  - A centralized buffer management to achieve enhanced buffer utilization.
  - Demonstrated a 50% decrease in total buffer requirement in their router.
  - Did not provide an active power management strategy.

Prior Research …

- Wang et al. DATE 2008:
  - Proposed a zero-efficient design for router buffers that optimizes the circuit level design of router buffer.
  - Basis of their research is predominance of zeros in the NoC traffic.
  - This is primarily a circuit level work under the assumption of high zero density and does not necessarily fare well when there is majority of one.
  - They do not consider any system-level information or active power management technique to adapt to the dynamic nature of the traffic.

Router Architecture

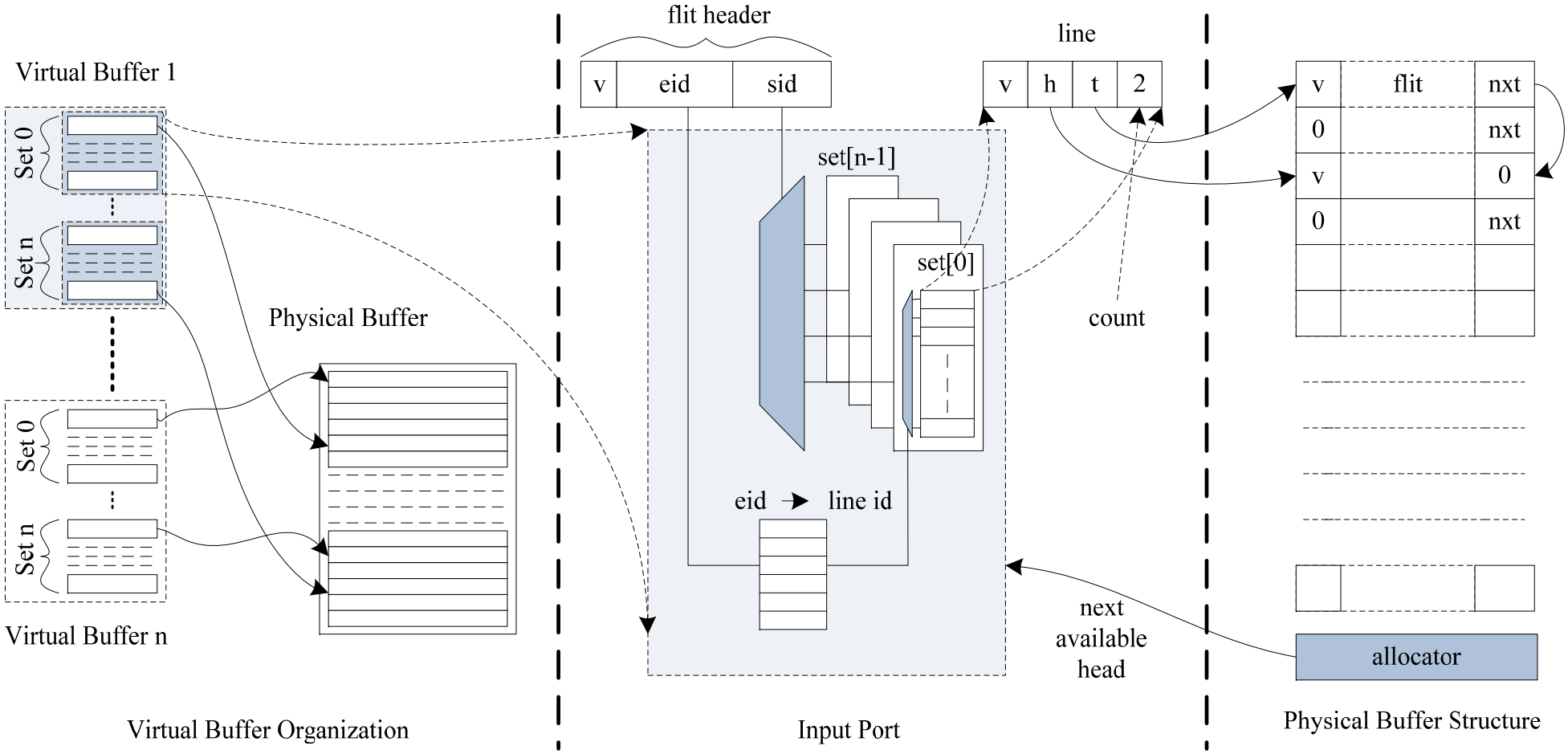
Router Architecture: Features

- The proposed buffer design is suitable for routers with centralized buffer management.
- To effectively utilize the central buffer design a concept of virtual buffer is introduced.
- Queue management is performed in the physical buffer.
- A concept of set and line is introduced for allocating buffer.

Router Architecture: Features ...

- The virtual buffers allow independent management of the central buffer structure.
- The physical buffer is managed centrally and each virtual buffer may or may not be mapped to a physical buffer.
- To be able to effectively perform power management using power gating the buffer is grouped in blocks.

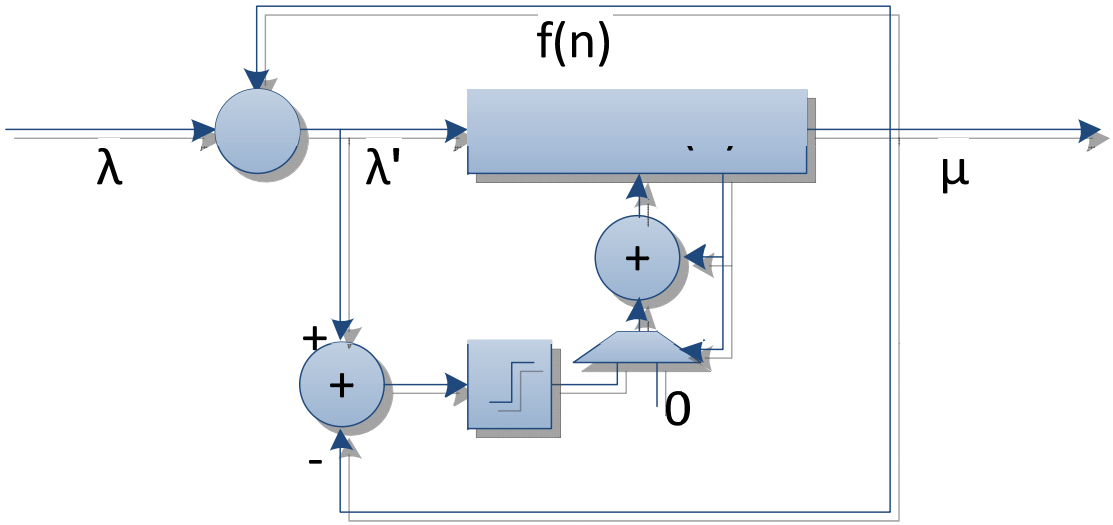
Central Buffer Router Architecture



Dynamic Buffer Management

Block-Level Power Management

- Traffic flow is modeled as a feedback loop.
- The buffer size is controlled by a threshold function.

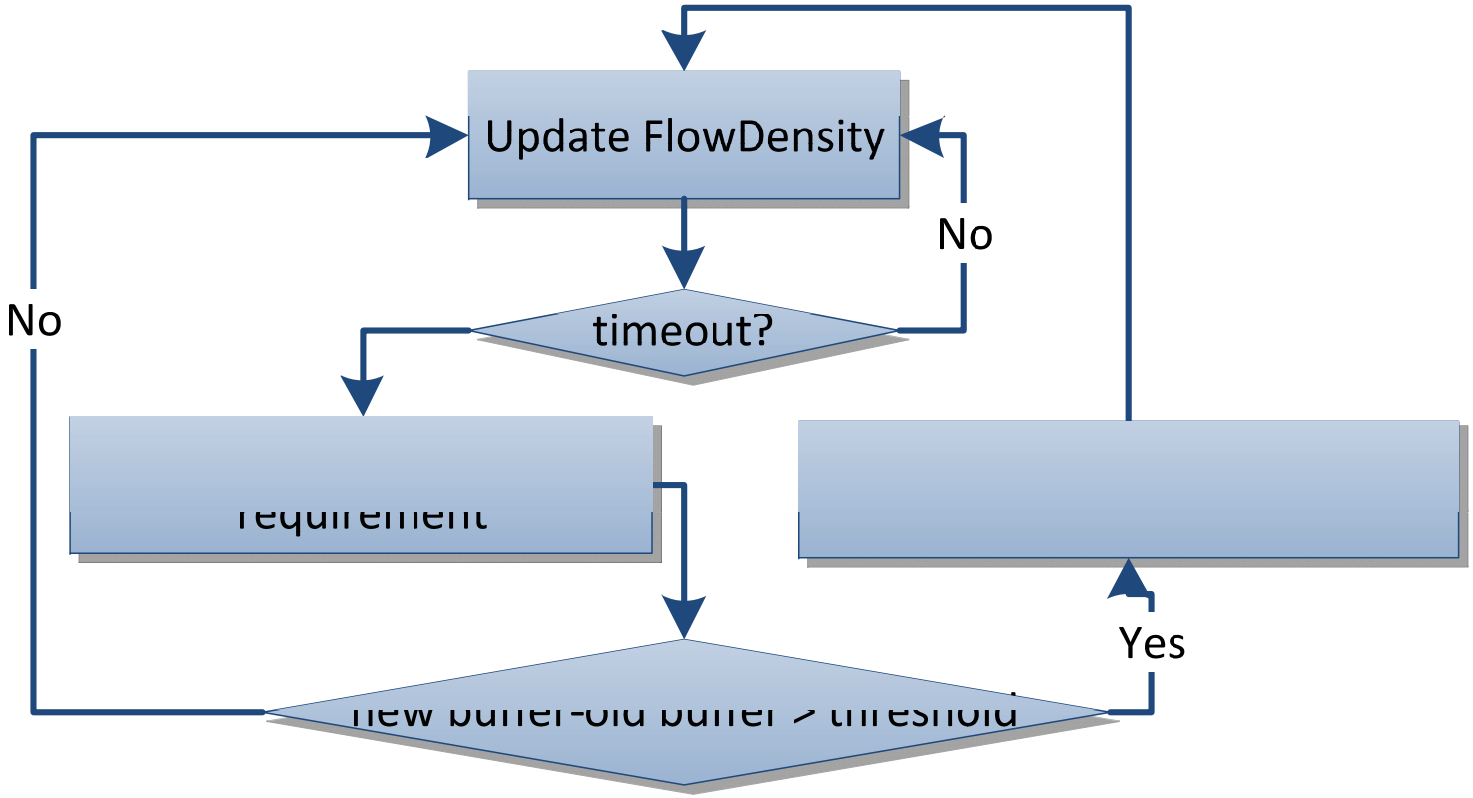


λ = Buffer Allocation Rate
μ = Buffer Free Rate

λ' = observed traffic
f = back pressure

Block-Level Feedback System

Controller FSM



How to Resize Buffer

Solution

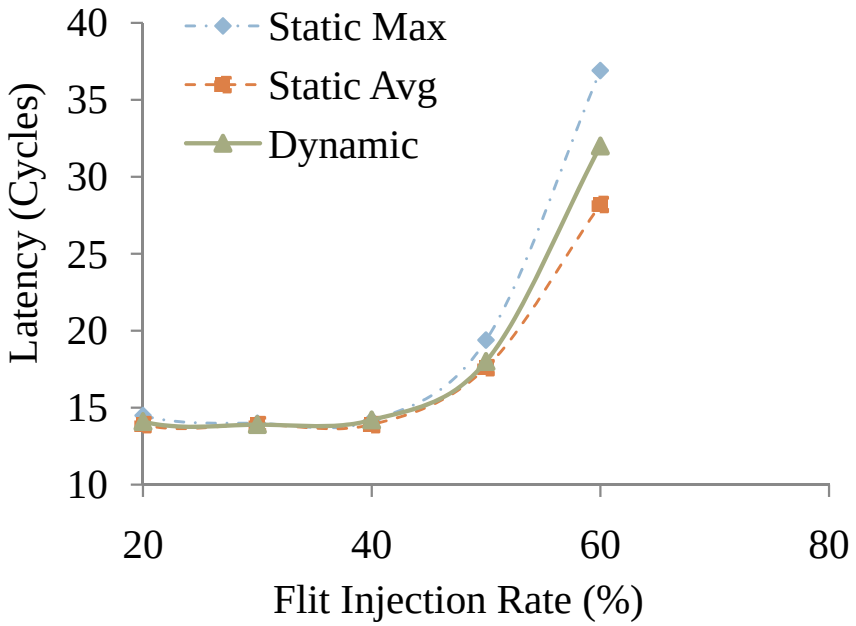
- Organize Buffer in blocks
- Turn of un-used blocks

Challenge

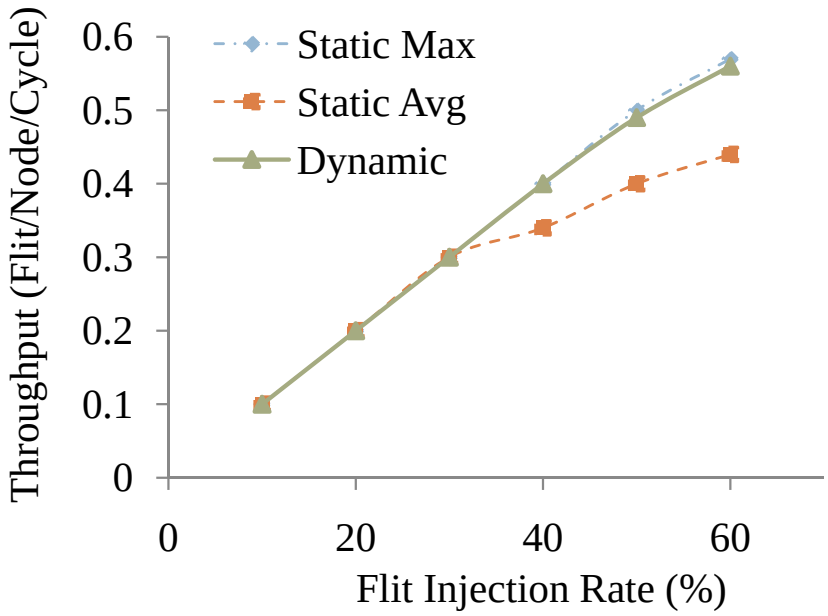
- Central Buffer Router is needed

Performance Results

Latency Comparison

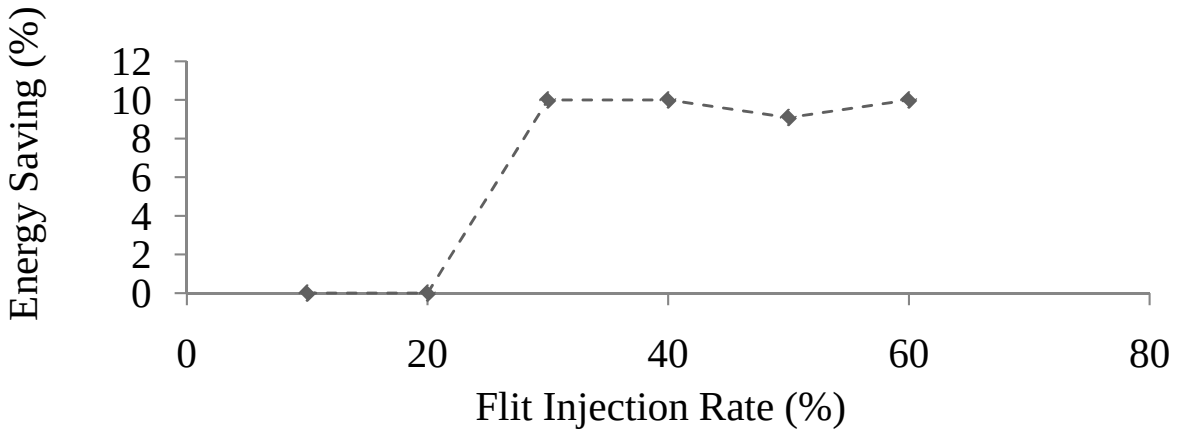


Throughput Comparison



21% Throughput Improvement at negligible loss of latency

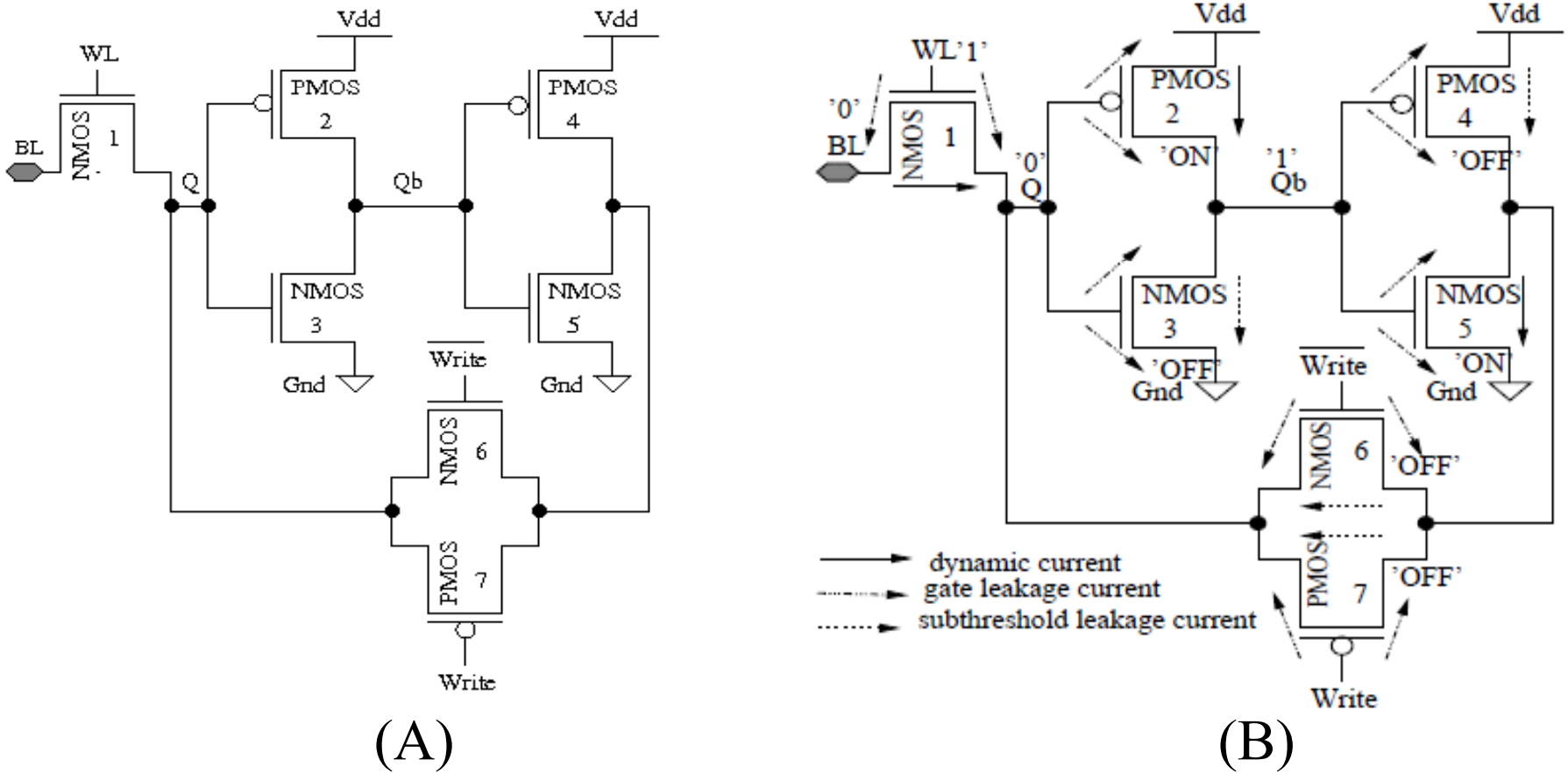
Energy Savings



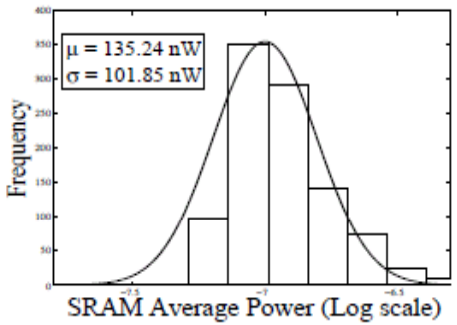
10% Energy Savings Compared to Static Buffer

Low Power SRAM Buffer

7-Transistor Low Leakage SRAM



SRAM Power Model



Total Power (7T – 45nm)	
Total Power	100.5 nW
Static Noise Margin	303.3 mV

0 and 1 does not require same energy!

Static and Dynamic Power Dissipation of SRAM.			
Power	Operation	Mean (μ)	Standard Deviation (σ)
Gate Leakage	Write 1	21.2nW	9.4nW
	Write 0	21.9nW	9.5nW
	Read 1	12.9nW	5.4nW
	Read 0	7.8nW	3.2nW
	Store 1	2.8nW	1.8nW
	Store 0	1.0nW	0.5nW
Subthreshold Leakage	Write 1	38.2nW	21.1nW
	Write 0	7.8nW	19.0nW
	Read 1	12.3nW	27.0nW
	Read 0	13.5nW	32.1nW
	Store 1	10.8nW	21.0nW
	Store 0	16.2nW	2.3nW
Dynamic Power	Write 1	39.2nW	22.1nW
	Write 0	5.1nW	20.0nW
	Read 1	14.3nW	30.0nW
	Read 0	15.5nW	32.1nW
	Store 1	12.8nW	22.0nW
	Store 0	17.2nW	2.9nW

Idea!

- Encode flits so that the storage is least energy consuming
- Done by utilizing system level information about flit content

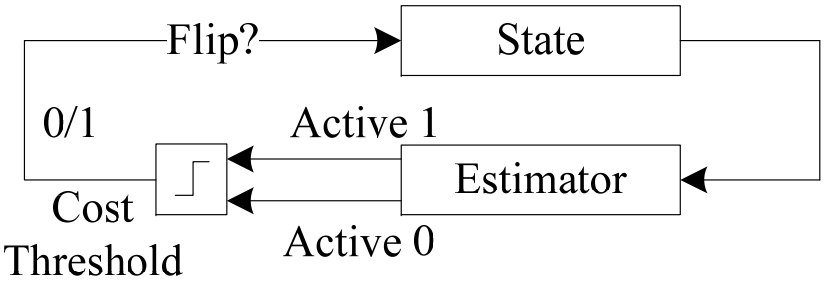
Flit-Level Power Management

Flit-Level Power Management: Approach

- A dynamic encoding technique is applied per flit to for further energy efficiency.
- Invert flits before writing to buffer if resulting energy consumption is less.
- Use an adaptive controller to trigger inversion.

The Adaptive Controller

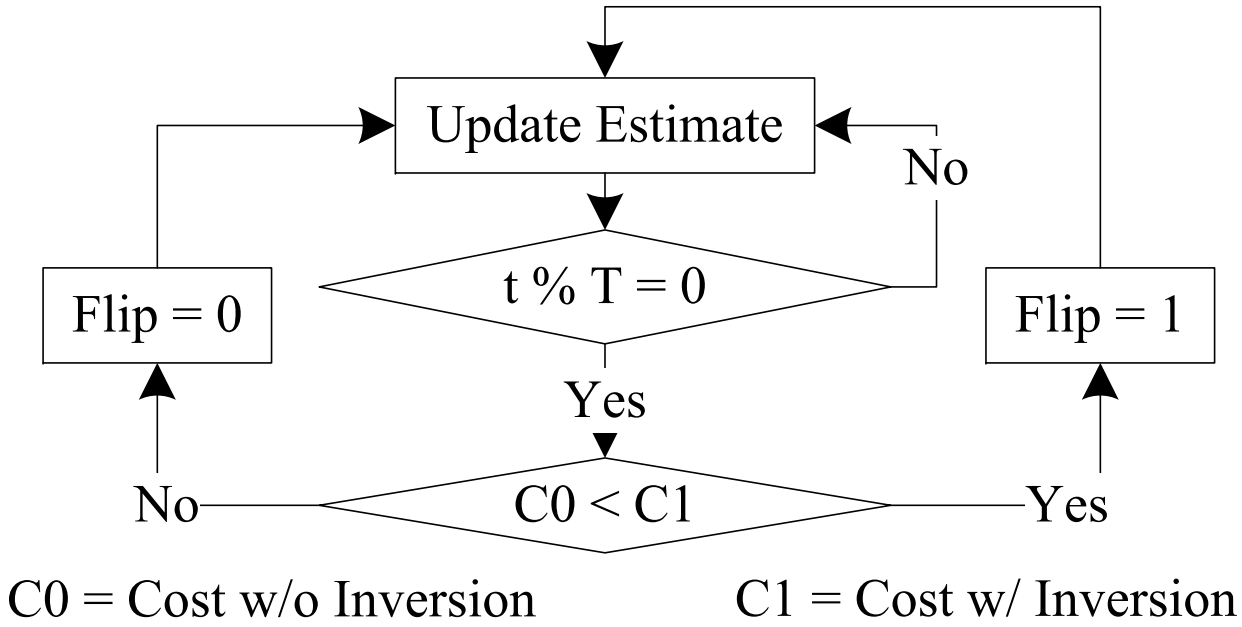
- Any flit can be stored in one of the three states: Active 0, Active 1 or Sleep.
- A linear adaptive control mechanism is designed to assign the flit storage states dynamically.



The Adaptive Controller for State Assignment

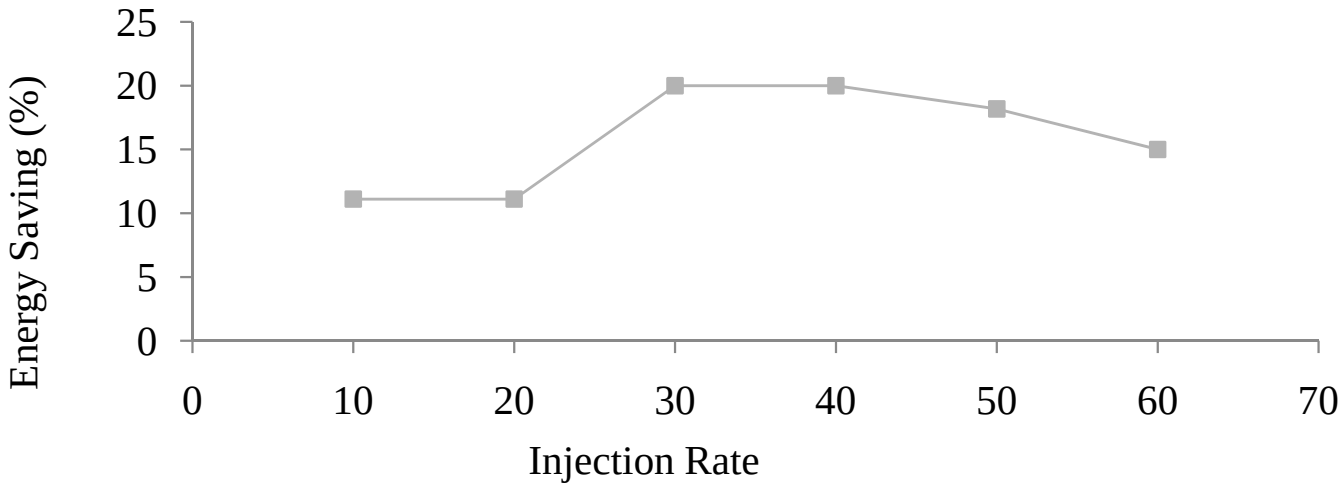
- A simple estimator is designed for low overhead.
- Flits are marked to be '1-dense' by adding a bit to header.
- A simple estimate is the frequency of this bit being set.

The Adaptive Controller



Controller FSM

Energy Savings



20% Energy Savings Compared to Generic Design

Conclusions

Conclusions

- A novel low power nano-CMOS buffer design was presented.
- Combined block and flit level power management is performed for throughput and power efficiency.
- Proposed technique utilizes system level information for effective power management of router buffer.
- Experimental evaluation have demonstrated the proposed design to be outperforming static buffer allocation by 21% in terms of throughput while consuming up to 20% less power.

Thank You!